DocCode – SCORE

SCORE Placeholder Sheet for IFW Content

Application Number: 16405612 Document Date: 05/07/2019

The presence of this form in the IFW record indicates that the following document type was received in electronic format on the date identified above. This content is stored in the SCORE database.

Since this was an electronic submission, there is no physical artifact folder, no artifact folder is recorded in PALM, and no paper documents or physical media exist. The TIFF images in the IFW record were created from the original documents that are stored in SCORE.

- Drawing

At the time of document entry (noted above):

- USPTO employees may access SCORE content via DAV or via the SCORE web page.
- External customers may access SCORE content via PAIR using the Supplemental Content tab.

Form Revision Date: March 1, 2019

CERTIFICATION OF MICRO ENTITY STATUS
(INSTITUTION OF HIGHER EDUCATION BASIS)

Application Number or Control Number (if applicable):	Patent Number (if applicable):
First Named Inventor: Revanth Patil	Title of Invention: SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ON

CERTIFICATION ON THE BASIS OF EMPLOYMENT BY AN INSTITUTION OF HIGHER EDUCATION

The applicant hereby certifies the following—

The applicant qualifies as a small entity as defined in 37 CFR 1.27.

The applicant's employer, from which the applicant obtains the majority of the applicant's income, is an institution of higher education as defined in section 101(a) of the Higher Education Act of 1965 (20 U.S.C. 1001(a)).

SIGNATURE by an authorized party set forth in 37 CFR 1.33(b)

Signature					
Name					
Date		Telephone		Registration No.	

☐ There is more than one inventor and I am one of the inventors who are jointly identified as the applicant. The required additional certification form(s) signed by the other joint inventor(s) are included with this form.

CERTIFICATION ON THE BASIS OF ASSIGNMENT OR OBLIGATION TO ASSIGN TO AN INSTITUTION OF HIGHER EDUCATION

The applicant hereby certifies the following—

The applicant qualifies as a small entity as defined in 37 CFR 1.27.

The applicant has assigned, granted, conveyed, or is under an obligation by contract or law, to assign, grant, or convey, a license or other ownership interest in the particular application to an institution of higher education as defined in section 101(a) of the Higher Education Act of 1965 (20 U.S.C. 1001(a)).

SIGNATURE by an authorized party set forth in 37 CFR 1.33(b)

Signature	/Timothy Fontes/				
Name	Timothy Fontes				
Date	2019-05-07	Telephone	602-650-2000	Registration No.	70801

☐ There is more than one inventor and I am one of the inventors who are jointly identified as the applicant. The required additional certification form(s) signed by the other joint inventor(s) are included with this form.

NOTE: A signed copy of this "institution of higher education basis" micro entity certification form (PTO/SB/15B) filed together with a signed copy of the "gross income basis" micro entity certification form (PTO/SB/15A) will be taken to be a representation that any inventor or person applying for a patent as provided in 37 CFR 1.46 not qualifying on the "gross income basis" qualifies on the "institution of higher education basis."

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Electronic Patent Application Fee Transmittal

Application Number:	
Filing Date:	
Title of Invention:	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB
First Named Inventor/Applicant Name:	Revanth Patil
Filer:	Timothy D. Fontes/Rebecca Endsley
Attorney Docket Number:	055743-621457 (M18-191P)

Filed as Micro Entity

Filing Fees for Utility under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
BASIC FILING FEE-UTILITY	3011	1	75	75
UTILITY SEARCH FEE	3111	1	165	165
UTILITY EXAMINATION FEE	3311	1	190	190
Pages:				
Claims:				
Miscellaneous-Filing:				
LATE FILING FEE FOR OATH OR DECLARATION	3051	1	40	40
Petition:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				
Miscellaneous:				
		Total in USD ($)		**470**

Electronic Acknowledgement Receipt

EFS ID:	35944420
Application Number:	16405612
International Application Number:	
Confirmation Number:	4562
Title of Invention:	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB
First Named Inventor/Applicant Name:	Revanth Patil
Customer Number:	27148
Filer:	Timothy D. Fontes/Rebecca Endsley
Filer Authorized By:	Timothy D. Fontes
Attorney Docket Number:	055743-621457 (M18-191P)
Receipt Date:	07-MAY-2019
Filing Date:	
Time Stamp:	17:12:22
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	DA
Payment was successfully received in RAM	$470
RAM confirmation Number	050819INTEFSW00004567501662
Deposit Account	
Authorized User	

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Application Data Sheet	WebADS.pdf	147340 10094e044dcfc5ad258f837e919025c1e4931767	no	8
Warnings:					
Information:					
2	Drawings-only black and white line drawings	621457_Figures.pdf	1371053 c79447e76925a236999a7ee52a201fff6e1bb0a6	no	11
Warnings:					
Information:					
3		621457_Specification_Claims_Abstract.pdf	373397 0000fed2486a157bbfd6ce4562128a4c9f686dee	yes	33

Multipart Description/PDF files in .zip description

Document Description	Start	End
Abstract	33	33
Claims	28	32
Specification	1	27

Warnings:					
Information:					
4	Certification of Micro Entity (Education Basis)	621457_Certification_of_Micro_Entity_Status.pdf	121434 38e00c6dbad3c97760966766c5aec7b783264af3	no	2
Warnings:					
Information:					
5	Fee Worksheet (SB06)	fee-info.pdf	37085 8ba871281e9a5f3f31d09b4e2ec48b9799941cea	no	2

Warnings:	
Information:	
Total Files Size (in bytes):	2050309

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	055743-621457 (M18-191P)
	Application Number	
Title of Invention	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB	

The application data sheet is part of the provisional or nonprovisional application for which it is being submitted. The following form contains the bibliographic data arranged in a format specified by the United States Patent and Trademark Office as outlined in 37 CFR 1.76.
This document may be completed electronically and submitted to the Office in electronic format using the Electronic Filing System (EFS) or the document may be printed and included in a paper filed application.

Secrecy Order 37 CFR 5.2:

☐ Portions or all of the application associated with this Application Data Sheet may fall under a Secrecy Order pursuant to 37 CFR 5.2 (Paper filers only. Applications that fall under Secrecy Order may not be filed electronically.)

Inventor Information:

Inventor 1

Legal Name

Prefix	Given Name	Middle Name	Family Name	Suffix
	Revanth		Patil	

Residence Information (Select One) ◉ US Residency ○ Non US Residency ○ Active US Military Service

City	Tempe	State/Province	AZ	Country of Residence i	US

Mailing Address of Inventor:

Address 1	699 S. Mills Avenue		
Address 2	Suite 601, Room 691AA		
City	Tempe	State/Province	AZ
Postal Code	85281	Country i	US

Inventor 2

Legal Name

Prefix	Given Name	Middle Name	Family Name	Suffix
	Paulo		Shakarian	

Residence Information (Select One) ◉ US Residency ○ Non US Residency ○ Active US Military Service

City	Tempe	State/Province	AZ	Country of Residence i	US

Mailing Address of Inventor:

Address 1	699 S. Mills Avenue		
Address 2	Suite 601, Room 691AA		
City	Tempe	State/Province	AZ
Postal Code	85281	Country i	US

Inventor 3

Legal Name

Prefix	Given Name	Middle Name	Family Name	Suffix
	Ashkan		Aleali	

IPR2025-01398
CrowdStrike EX1002 Page 9

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	055743-621457 (M18-191P)
	Application Number	
Title of Invention	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB	

Residence Information (Select One)	◉ US Residency	○ Non US Residency	○ Active US Military Service		
City	Tempe	**State/Province**	AZ	**Country of Residence** i	US

Mailing Address of Inventor:

Address 1	699 S. Mills Avenue		
Address 2	Suite 601, Room 691AA		
City	Tempe	**State/Province**	AZ
Postal Code	85281	**Country** i	US

Inventor 4

Legal Name

Prefix	**Given Name**	**Middle Name**	**Family Name**	**Suffix**
	Ericsson		Marin	

Residence Information (Select One)	◉ US Residency	○ Non US Residency	○ Active US Military Service		
City	Tempe	**State/Province**	AZ	**Country of Residence** i	US

Mailing Address of Inventor:

Address 1	699 S. Mills Avenue		
Address 2	Suite 601, Room 691AA		
City	Tempe	**State/Province**	AZ
Postal Code	85281	**Country** i	US

All Inventors Must Be Listed - Additional Inventor Information blocks may be generated within this form by selecting the **Add** button. Add

Correspondence Information:

Enter either Customer Number or complete the Correspondence Information section below.
For further information see 37 CFR 1.33(a).

☐ **An Address is being provided for the correspondence Information of this application.**

Customer Number	27148		
Email Address	uspt@polsinelli.com	Add Email	Remove Email

Application Information:

Title of the Invention	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB		
Attorney Docket Number	055743-621457 (M18-191P)	**Small Entity Status Claimed**	☒
Application Type	Nonprovisional		
Subject Matter	Utility		
Total Number of Drawing Sheets (if any)	11	**Suggested Figure for Publication (if any)**	

PTO/AIA/14 (08-15)
Approved for use through 04/30/2017. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	055743-621457 (M18-191P)
	Application Number	
Title of Invention	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB	

Filing By Reference:

Only complete this section when filing an application by reference under 35 U.S.C. 111(c) and 37 CFR 1.57(a). Do not complete this section if application papers including a specification and any drawings are being filed. Any domestic benefit or foreign priority information must be provided in the appropriate section(s) below (i.e., "Domestic Benefit/National Stage Information" and "Foreign Priority Information").

For the purposes of a filing date under 37 CFR 1.53(b), the description and any drawings of the present application are replaced by this reference to the previously filed application, subject to conditions and requirements of 37 CFR 1.57(a).

Application number of the previously filed application	Filing date (YYYY-MM-DD)	Intellectual Property Authority or Country

Publication Information:

☐ Request Early Publication (Fee required at time of Request 37 CFR 1.219)

☐ **Request Not to Publish.** I hereby request that the attached application not be published under 35 U.S.C. 122(b) and certify that the invention disclosed in the attached application **has not and will not** be the subject of an application filed in another country, or under a multilateral international agreement, that requires publication at eighteen months after filing.

Representative Information:

Representative information should be provided for all practitioners having a power of attorney in the application. Providing this information in the Application Data Sheet does not constitute a power of attorney in the application (see 37 CFR 1.32).
Either enter Customer Number or complete the Representative Name section below. If both sections are completed the customer Number will be used for the Representative Information during processing.

Please Select One:	◉ Customer Number	○ US Patent Practitioner	○ Limited Recognition (37 CFR 11.9)
Customer Number	27148		

Prefix	**Given Name**	Middle Name	**Family Name**	Suffix	Remove
Registration Number					
Prefix	**Given Name**	Middle Name	**Family Name**	Suffix	Remove
Registration Number					

Additional Representative Information blocks may be generated within this form by selecting the **Add** button.

PTO/AIA/14 (08-15)
Approved for use through 04/30/2017. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	055743-621457 (M18-191P)
	Application Number	
Title of Invention	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB	

Domestic Benefit/National Stage Information:

This section allows for the applicant to either claim benefit under 35 U.S.C. 119(e), 120, 121, 365(c), or 386(c) or indicate National Stage entry from a PCT application. Providing benefit claim information in the Application Data Sheet constitutes the specific reference required by 35 U.S.C. 119(e) or 120, and 37 CFR 1.78.
When referring to the current application, please leave the "Application Number" field blank.

Prior Application Status	Pending		Remove
Application Number	Continuity Type	Prior Application Number	Filing or 371(c) Date (YYYY-MM-DD)
	Claims benefit of provisional	62668878	2018-05-09

Additional Domestic Benefit/National Stage Data may be generated within this form by selecting the **Add** button.

Foreign Priority Information:

This section allows for the applicant to claim priority to a foreign application. Providing this information in the application data sheet constitutes the claim for priority as required by 35 U.S.C. 119(b) and 37 CFR 1.55. When priority is claimed to a foreign application that is eligible for retrieval under the priority document exchange program (PDX) the information will be used by the Office to automatically attempt retrieval pursuant to 37 CFR 1.55(i)(1) and (2). Under the PDX program, applicant bears the ultimate responsibility for ensuring that a copy of the foreign application is received by the Office from the participating foreign intellectual property office, or a certified copy of the foreign priority application is filed, within the time period specified in 37 CFR 1.55(g)(1).

Application Number	Country	Filing Date (YYYY-MM-DD)	Access Code (if applicable)

Additional Foreign Priority Data may be generated within this form by selecting the **Add** button.

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications

☐ This application (1) claims priority to or the benefit of an application filed before March 16, 2013 and (2) also contains, or contained at any time, a claim to a claimed invention that has an effective filing date on or after March 16, 2013.
NOTE: By providing this statement under 37 CFR 1.55 or 1.78, this application, with a filing date on or after March 16, 2013, will be examined under the first inventor to file provisions of the AIA.

PTO/AIA/14 (08-15)
Approved for use through 04/30/2017. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	055743-621457 (M18-191P)
	Application Number	
Title of Invention	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB	

Authorization or Opt-Out of Authorization to Permit Access:

When this Application Data Sheet is properly signed and filed with the application, applicant has provided written authority to permit a participating foreign intellectual property (IP) office access to the instant application-as-filed (see paragraph A in subsection 1 below) and the European Patent Office (EPO) access to any search results from the instant application (see paragraph B in subsection 1 below).

Should applicant choose not to provide an authorization identified in subsection 1 below, applicant **must opt-out** of the authorization by checking the corresponding box A or B or both in subsection 2 below.

NOTE: This section of the Application Data Sheet is **ONLY** reviewed and processed with the **INITIAL** filing of an application. After the initial filing of an application, an Application Data Sheet cannot be used to provide or rescind authorization for access by a foreign IP office(s). Instead, Form PTO/SB/39 or PTO/SB/69 must be used as appropriate.

1. Authorization to Permit Access by a Foreign Intellectual Property Office(s)

A. Priority Document Exchange (PDX) - Unless box A in subsection 2 (opt-out of authorization) is checked, the undersigned hereby **grants the USPTO authority** to provide the European Patent Office (EPO), the Japan Patent Office (JPO), the Korean Intellectual Property Office (KIPO), the State Intellectual Property Office of the People's Republic of China (SIPO), the World Intellectual Property Organization (WIPO), and any other foreign intellectual property office participating with the USPTO in a bilateral or multilateral priority document exchange agreement in which a foreign application claiming priority to the instant patent application is filed, access to: (1) the instant patent application-as-filed and its related bibliographic data, (2) any foreign or domestic application to which priority or benefit is claimed by the instant application and its related bibliographic data, and (3) the date of filing of this Authorization. See 37 CFR 1.14(h)(1).

B. Search Results from U.S. Application to EPO - Unless box B in subsection 2 (opt-out of authorization) is checked, the undersigned hereby **grants the USPTO authority** to provide the EPO access to the bibliographic data and search results from the instant patent application when a European patent application claiming priority to the instant patent application is filed. See 37 CFR 1.14(h)(2).

The applicant is reminded that the EPO's Rule 141(1) EPC (European Patent Convention) requires applicants to submit a copy of search results from the instant application without delay in a European patent application that claims priority to the instant application.

2. Opt-Out of Authorizations to Permit Access by a Foreign Intellectual Property Office(s)

☐ A. Applicant **DOES NOT** authorize the USPTO to permit a participating foreign IP office access to the instant application-as-filed. If this box is checked, the USPTO will not be providing a participating foreign IP office with any documents and information identified in subsection 1A above.

☐ B. Applicant **DOES NOT** authorize the USPTO to transmit to the EPO any search results from the instant patent application. If this box is checked, the USPTO will not be providing the EPO with search results from the instant application.

NOTE: Once the application has published or is otherwise publicly available, the USPTO may provide access to the application in accordance with 37 CFR 1.14.

PTO/AIA/14 (08-15)
Approved for use through 04/30/2017. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	055743-621457 (M18-191P)
	Application Number	
Title of Invention	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB	

Applicant Information:

Providing assignment information in this section does not substitute for compliance with any requirement of part 3 of Title 37 of CFR to have an assignment recorded by the Office.

Applicant 1

If the applicant is the inventor (or the remaining joint inventor or inventors under 37 CFR 1.45), this section should not be completed. The information to be provided in this section is the name and address of the legal representative who is the applicant under 37 CFR 1.43; or the name and address of the assignee, person to whom the inventor is under an obligation to assign the invention, or person who otherwise shows sufficient proprietary interest in the matter who is the applicant under 37 CFR 1.46. If the applicant is an applicant under 37 CFR 1.46 (assignee, person to whom the inventor is obligated to assign, or person who otherwise shows sufficient proprietary interest) together with one or more joint inventors, then the joint inventor or inventors who are also the applicant should be identified in this section.

Clear

◯ Assignee	◯ Legal Representative under 35 U.S.C. 117	◯ Joint Inventor
◯ Person to whom the inventor is obligated to assign.	◯ Person who shows sufficient proprietary interest	

If applicant is the legal representative, indicate the authority to file the patent application, the inventor is:

Name of the Deceased or Legally Incapacitated Inventor:	
If the Applicant is an Organization check here. ☐	

Prefix	**Given Name**	Middle Name	**Family Name**	Suffix

Mailing Address Information For Applicant:

Address 1			
Address 2			
City		**State/Province**	
Country[i]		Postal Code	
Phone Number		Fax Number	
Email Address			

Additional Applicant Data may be generated within this form by selecting the Add button.

PTO/AIA/14 (08-15)
Approved for use through 04/30/2017. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	055743-621457 (M18-191P)
	Application Number	
Title of Invention	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB	

Assignee Information including Non-Applicant Assignee Information:

Providing assignment information in this section does not substitute for compliance with any requirement of part 3 of Title 37 of CFR to have an assignment recorded by the Office.

Assignee 1

Complete this section if assignee information, including non-applicant assignee information, is desired to be included on the patent application publication. An assignee-applicant identified in the "Applicant Information" section will appear on the patent application publication as an applicant. For an assignee-applicant, complete this section only if identification as an assignee is also desired on the patent application publication.

If the Assignee or Non-Applicant Assignee is an Organization check here. ☒

Organization Name	Arizona Board of Regents on Behalf of Arizona State University

Mailing Address Information For Assignee including Non-Applicant Assignee:

Address 1	699 S. Mills Avenue		
Address 2	Suite 601, Room 691AA		
City	Tempe	**State/Province**	AZ
Country i	US	Postal Code	85281
Phone Number		Fax Number	
Email Address			

Additional Assignee or Non-Applicant Assignee Data may be generated within this form by selecting the Add button.

PTO/AIA/14 (08-15)
Approved for use through 04/30/2017. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	055743-621457 (M18-191P)
	Application Number	
Title of Invention	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB	

Signature:

NOTE: This Application Data Sheet must be signed in accordance with 37 CFR 1.33(b). **However, if this Application Data Sheet is submitted with the INITIAL filing of the application and either box A or B is not checked in subsection 2 of the "Authorization or Opt-Out of Authorization to Permit Access" section, then this form must also be signed in accordance with 37 CFR 1.14(c).**

This Application Data Sheet **must** be signed by a patent practitioner if one or more of the applicants is a **juristic entity** (e. g., corporation or association). If the applicant is two or more joint inventors, this form must be signed by a patent practitioner, **all** joint inventors who are the applicant, or one or more joint inventor-applicants who have been given power of attorney (e.g., see USPTO Form PTO/AIA/81) on behalf of **all** joint inventor-applicants.

See 37 CFR 1.4(d) for the manner of making signatures and certifications.

Signature	/Timothy Fontes/			Date (YYYY-MM-DD)	
First Name	Timothy	Last Name	Fontes	Registration Number	70801

Additional Signature may be generated within this form by selecting the Add button.

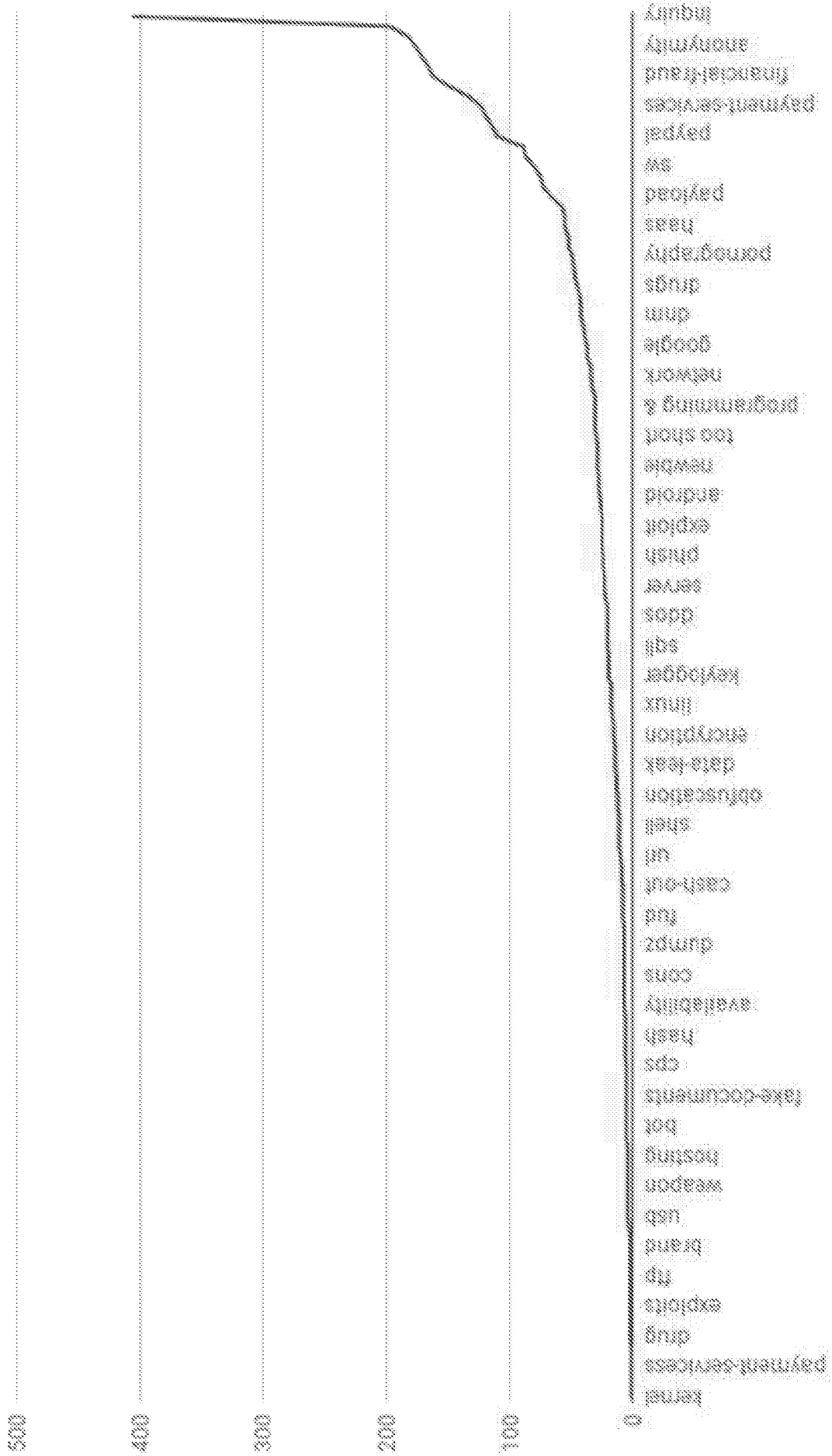

500
400
300
200
100
0
kernel
payment-services
drug
exploits
ftp
brand
usb
weapon
hosting
bot
take-documents
cps
heap
availability
coins
dumpz
fud
cash-out
url
shell
obfuscation
data-leak
encryption
linux
keylogger
sql
ddos
server
phish
exploit
android
newbie
too short
programming &
network
google
dnm
drugs
pornography
haas
payload
sw
paypal
payment-services
financial-fraud
anonymity
inquiry

FIG. 1

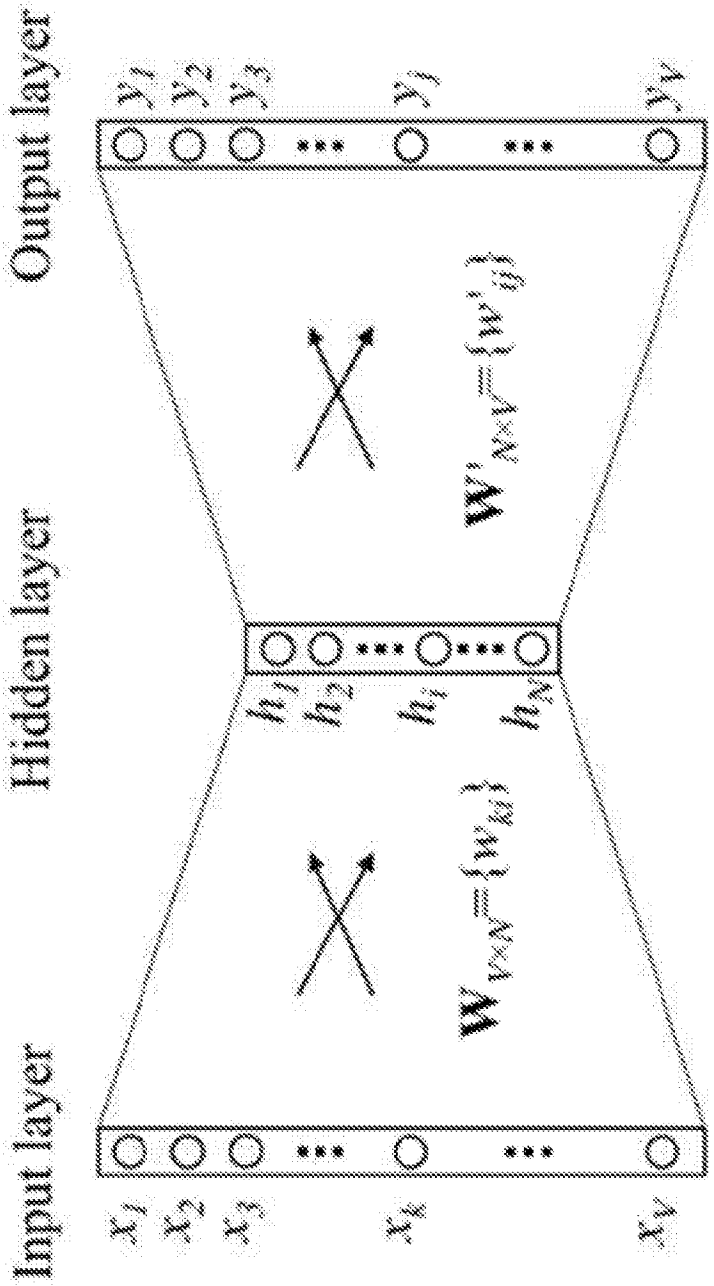

Output layer
Hidden layer
Input layer
y_1 y_2 y_3 y_j y_V
h_1 h_2 h_i h_N
x_1 x_2 x_3 x_k x_V
$\mathbf{W}'_{N\times V}=\{w'_{ij}\}$
$\mathbf{W}_{V\times N}=\{w_{ki}\}$

FIG. 2

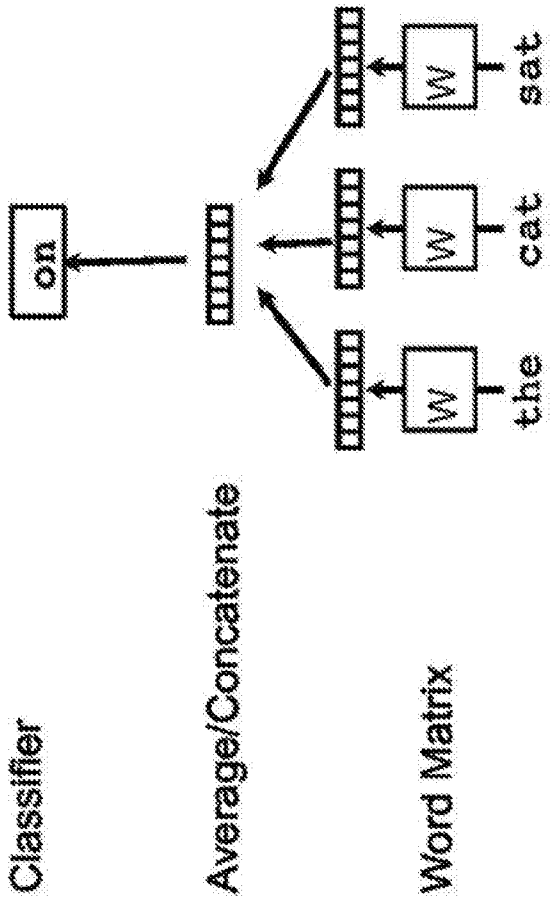

Classifier
Average/Concatenate
Word Matrix
on
the
cat
sat
W
W
W

FIG. 3

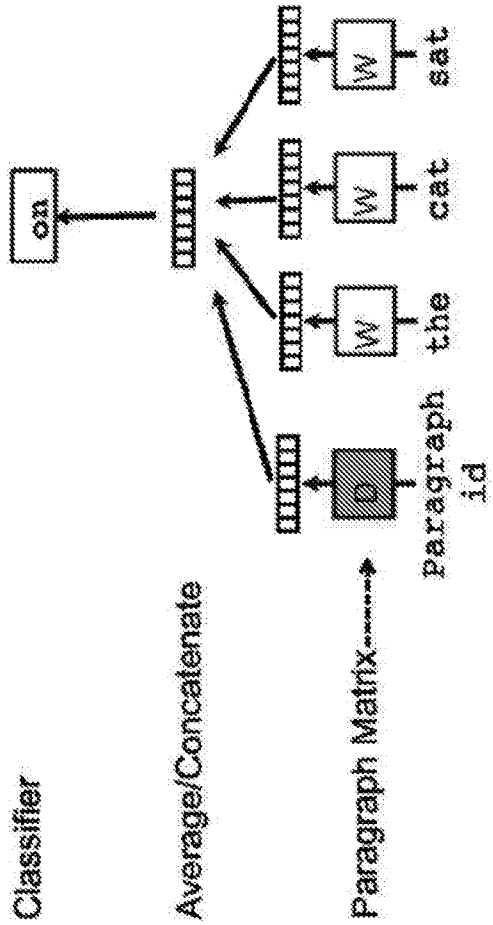

Classifier
Average/Concatenate
Paragraph Matrix
Paragraph id
the
cat
sat
on

FIG. 4

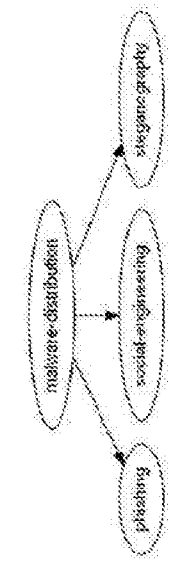


FIG. 5A

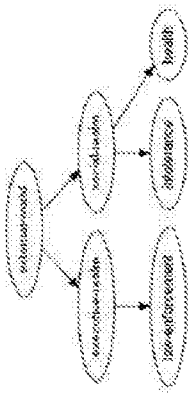


FIG. 5B

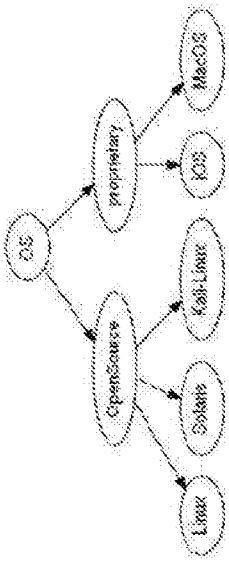


FIG. 5C

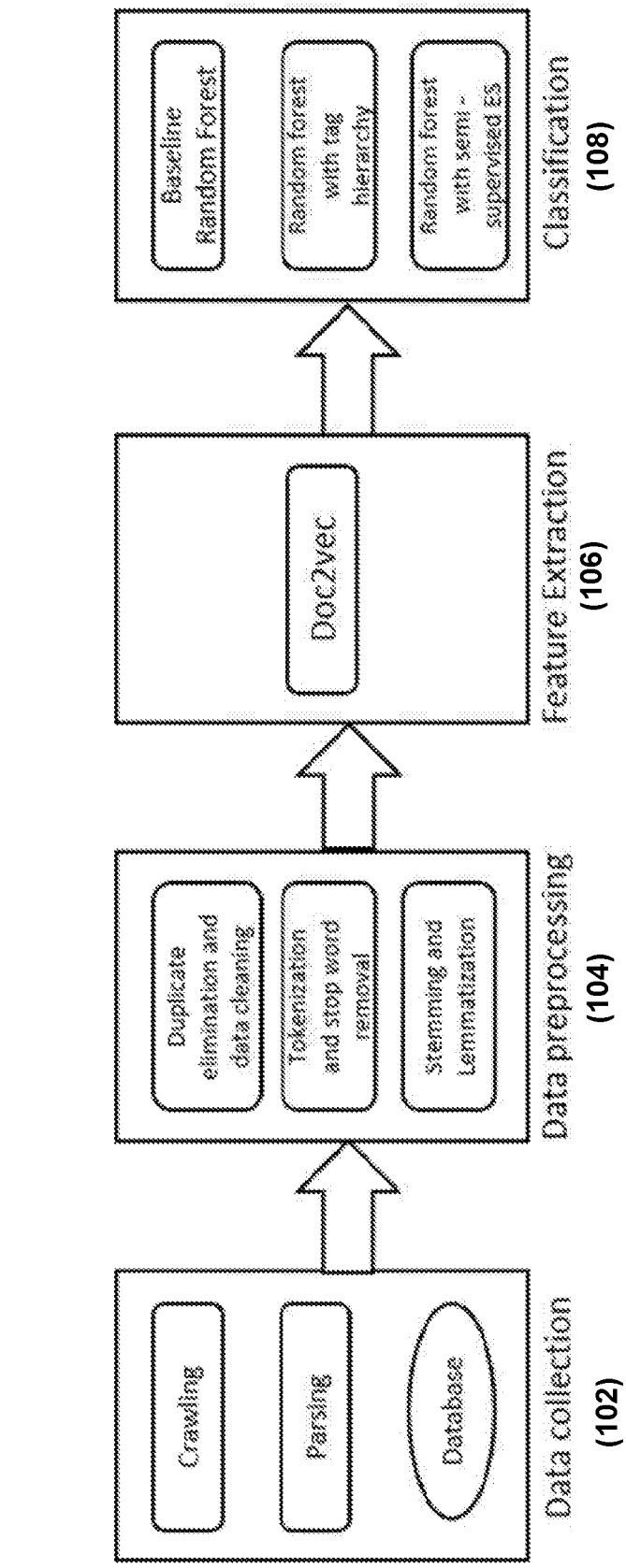

100
Data collection
(102)
Crawling
Parsing
Database
Data preprocessing
(104)
Duplicate elimination and data cleaning
Tokenization and stop word removal
Stemming and Lemmatization
Feature Extraction
(106)
Doc2vec
Classification
(108)
Baseline Random Forest
Random forest with tag hierarchy
Random forest with semi - supervised ES

FIG. 6

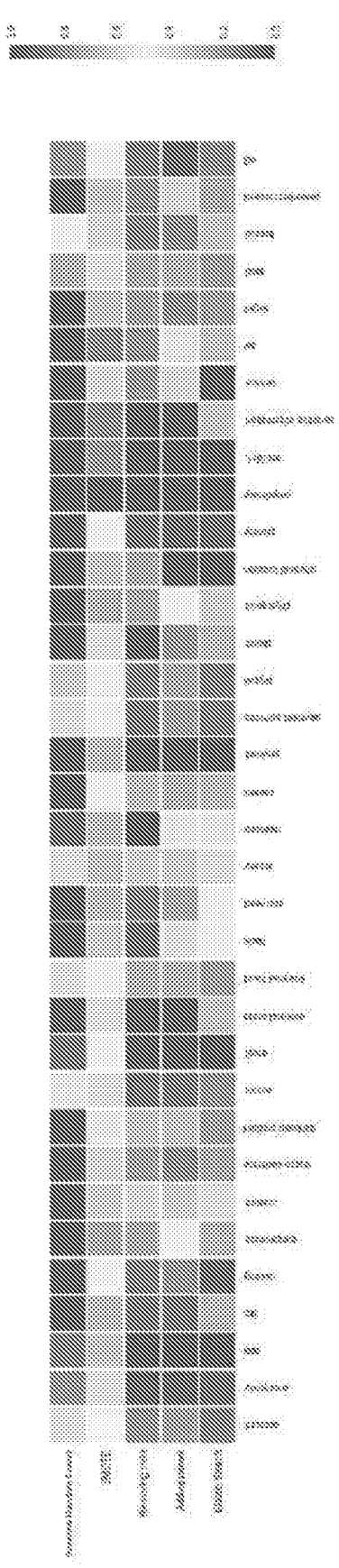


FIG. 7

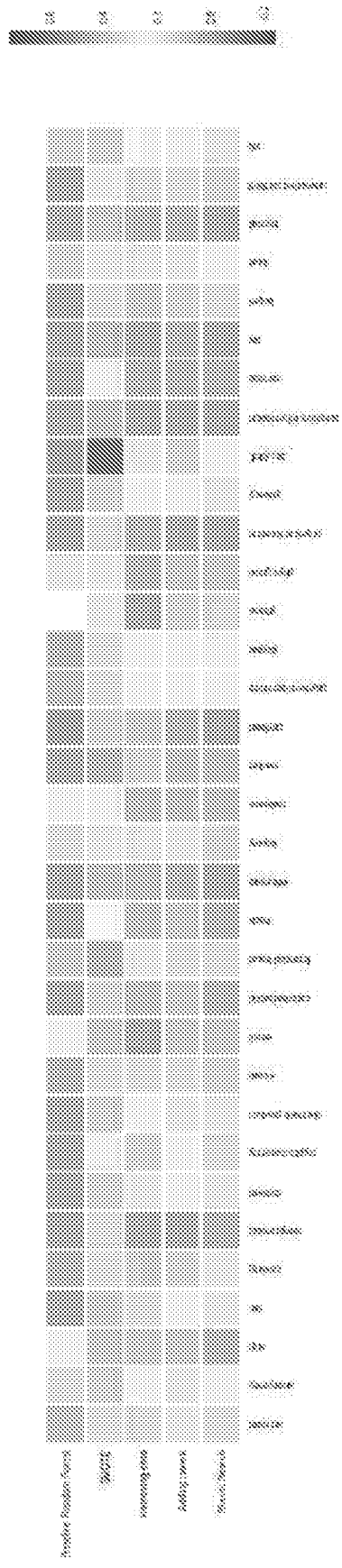


FIG. 8

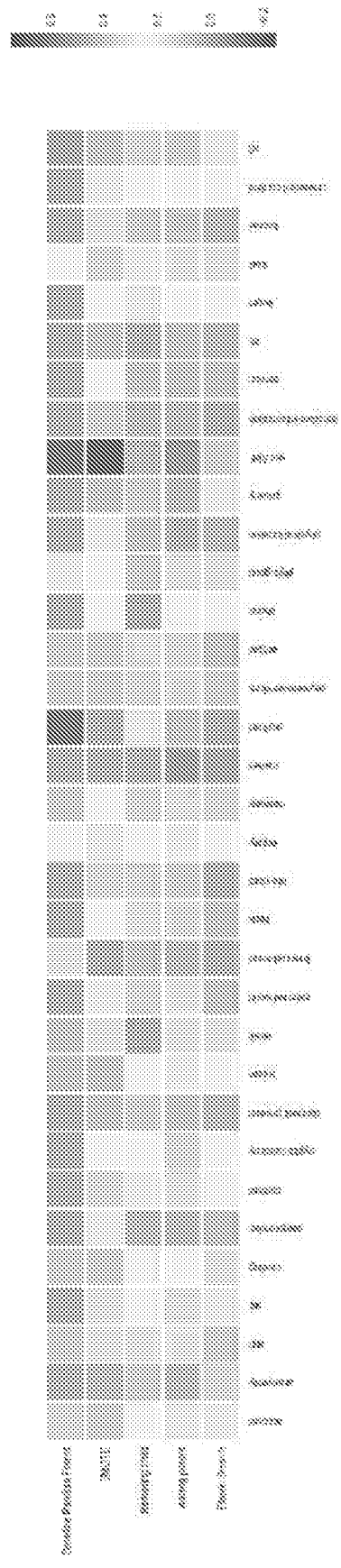


FIG. 9

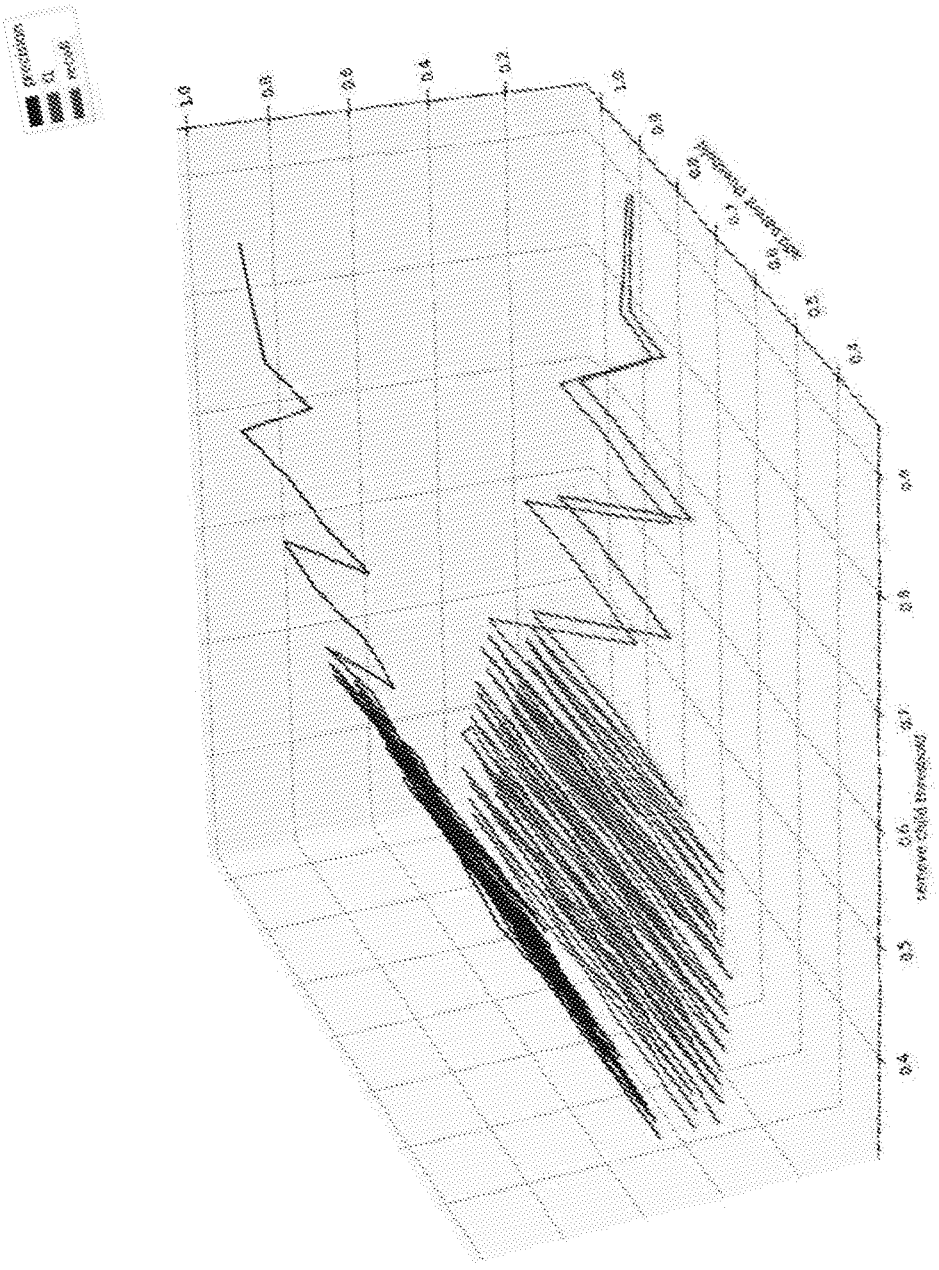


FIG. 10

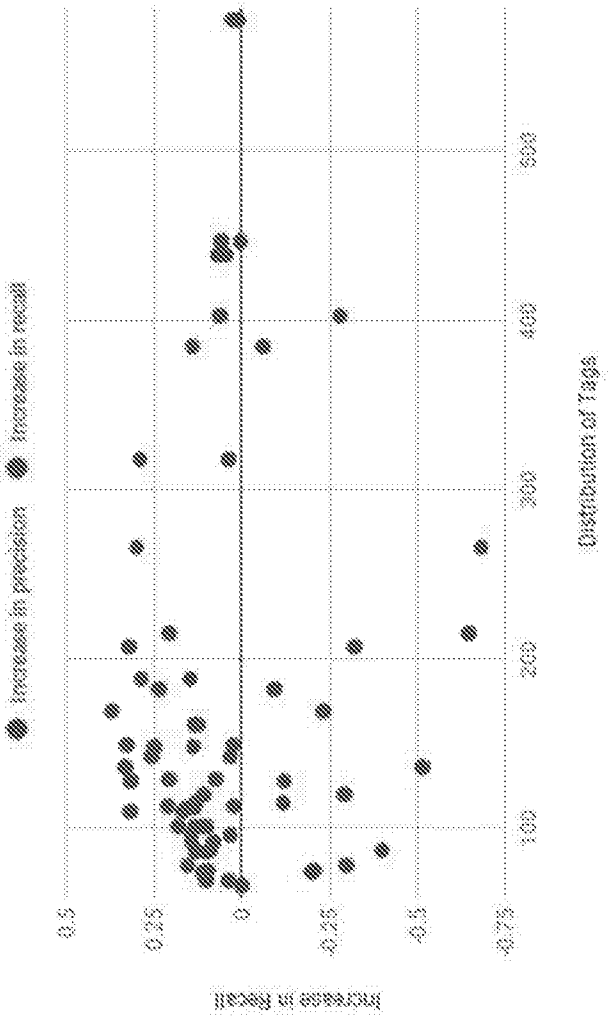

Increase in precision
Increase in recall
0.5
0.25
0
-0.25
-0.5
-0.75
100
200
300
400
500
Increase in f-score
Distribution of Tags

FIG. 11

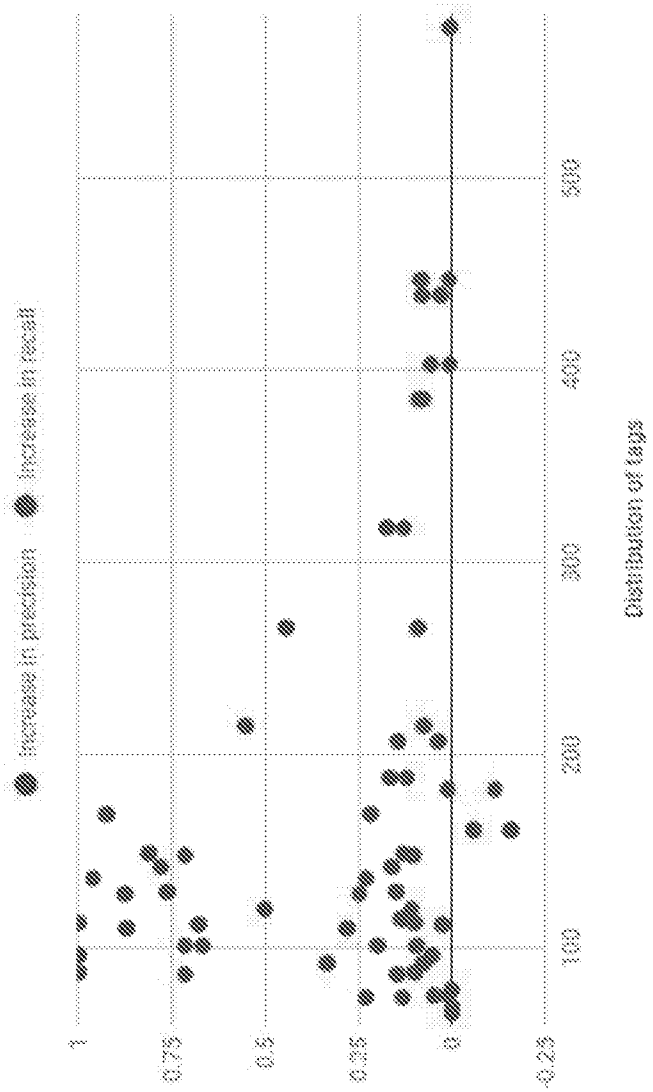

Increase in precision
Increase in recall
1
0.75
0.5
0.25
0
-0.25
100
200
300
400
500
Distribution of tags

FIG. 12

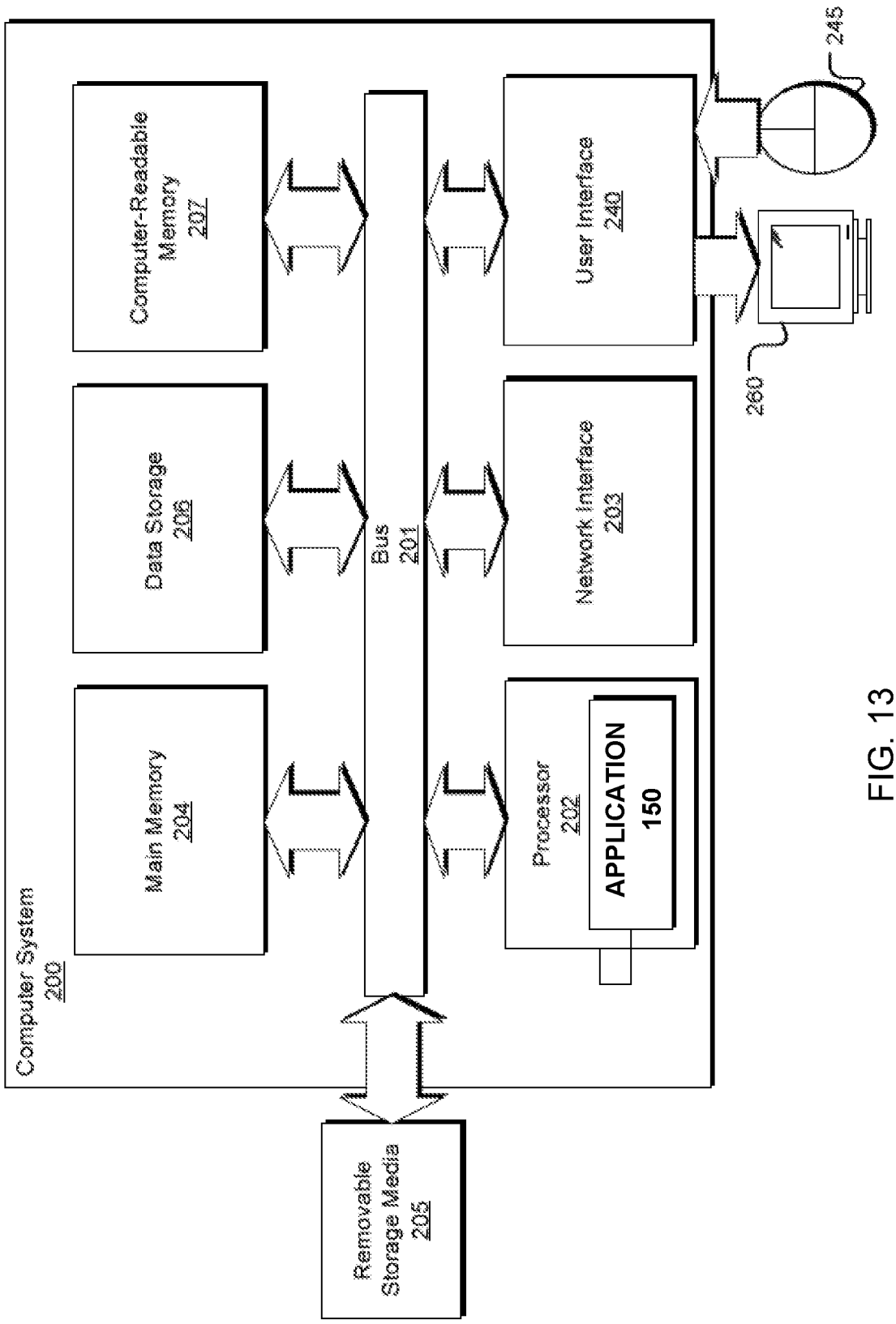

Computer System
200
Main Memory
204
Data Storage
206
Computer-Readable
Memory
207
Bus
201
Processor
202
APPLICATION
150
Network Interface
203
User Interface
240
245
260
Removable
Storage Media
205

FIG. 13

ABSTRACT

Embodiments of a computer-implemented system for improving classification of data associated with the deep web or dark net are disclosed.

CLAIMS

What is claimed is:

1. A computer-implemented system for improving classification of criminal activities arising from cyber environments, comprising:
   a processor configured to:
      access a ground truth dataset generated from deep web forum information, the ground truth dataset defining a predetermined tag hierarchy including predetermined tag labels corresponding topics of the deep web forum information;
      access data related to a discussion topic for classification from a deep web forum;
      extract a set of features from the data by assigning word vectors or paragraph vectors to the data; and
      apply a machine classifier to the set of features to generate a prediction list of tags for classifying the discussion topic,
      wherein the machine classifier is trained to maintain the predetermined tag hierarchy by regulating modification to parent and child tags associated with a predicted tag of the prediction list of tags based on a prediction probability of the predicted tag and predetermined tag thresholds assigned to the parent and child tags.

2. The computer-implemented system of claim 1, wherein the predetermined tag hierarchy is divided into parent tags and child tags, wherein a topic represented by a child tag is representative of a subset of a topic represented by a parent tag, wherein each child tag corresponds to a particular parent tag, and each tag is assigned a corresponding probability value.

3. The computer-implemented system of claim 1, wherein the processor is configured to remove a child tag of the prediction list of tags corresponding to a parent tag from the prediction list of tags where a probability value

associated with the child tag is below a predetermined threshold value.

4. The computer-implemented system of claim 1, wherein the processor is configured to add a parent tag corresponding to a child tag if a probability value associated with the parent tag is above a predetermined threshold value.

5. The computer-implemented system of claim 1, wherein the set of features are extracted using textual feature extraction to identify tags associated with topic titles of the data.

6. The computer-implemented system of claim 1, wherein a set of synthetic data samples are created for addition to the ground truth dataset to address class imbalance by creating a set of synthetic sample points along a line defined between a first minority feature vector and a second feature vector and adding the set of synthetic sample points to the ground truth dataset.

7. The computer-implemented system of claim 1, wherein the ground truth dataset is supplemented by, for every sample of the ground truth dataset in a minority class, adding top similar documents from the deep web forum information, as identified using an elastic search similarity score, and adding them to the ground truth dataset.

8. The computer-implemented system of claim 1, wherein the data is parsed and segmented into tokens, and predetermined characters and predetermined common words are removed from the data.

9. The computer-implemented system of claim 1, wherein words within the data are reduced to their root form using stemming and lemmatization procedures.

10. The computer-implemented system of claim 1, wherein the set of features are extracted from the data using a neural network, wherein the neural network is configured to learn a continuous distributed vector representation for documents and words within the data.

11. The computer-implemented system of claim 1, further comprising:

a crawler in operable communication with the processor and operable to traverse a plurality of deep web websites to retrieve Internet documents forming at least a portion of the deep web forum information.

12. A method, comprising:

configuring a processor for executing operations including:

accessing a plurality of documents from deep web forums, each of the plurality of documents defining a topic;

generating a ground truth dataset from the data, the ground truth dataset defining a predetermined tag hierarchy including predetermined tag labels corresponding to the topic of each of the plurality of documents;

modifying the ground truth dataset in view of a minority class to address class imbalance; and

generating a machine classifier to classify additional documents from the deep web forums, the machine classifier configured to maintain the predetermined tag hierarchy defined by the ground truth dataset.

13. The method of claim 12, further comprising modifying the ground truth dataset by introducing an additional document topically similar to the minority class, the additional document found by assigning an elastic search similarity score to each of the plurality of documents.

14. The method of claim 12, wherein a line is defined between a minority class feature vector and a nearest neighbor feature vector.

15. The method of claim 14, wherein a synthetic minority sample is added to the ground truth dataset by selecting a random point along the line defined between the minority class feature vector and the nearest neighbor feature vector is chosen and adding the random point to the minority class feature

vector.

16. The method of claim 12, further comprising generating the ground truth dataset by leveraging a spider program to crawl one or more deep web forums and retrieve HTML pages.

17. The method of claim 12, further comprising parsing the data to extract predetermined information related to the deep web forums including a topic title, a topic content, a post content, a topic author, a post author, an author reputation, and a posted date.

18. A method, comprising:

configuring a processor for executing operations including:

accessing data associated with a deep web forum, the data defining a topic for classification;

extracting a set of features from the data as inputs for a machine classifier; and

apply a machine classifier to the set of features to generate a prediction list of tags for classifying the topic,

wherein the machine classifier is configured to maintain a predetermined tag hierarchy by regulating modification to parent and child tags associated with a predicted tag of the prediction list of tags based on a prediction probability of the predicted tag and predetermined tag thresholds assigned to the parent and child tags.

19. The method of claim 18, further comprising:

outputting an update to the prediction list of tags, a total number of the tags of the prediction list of tags being different from the update to the prediction list of tags to maintain the predetermined tag hierarchy.

20. The method of claim 18, further comprising generating the predetermined tag hierarchy based on a ground truth dataset, and leveraging semi-supervised learning to address class imbalance and increase accuracy of the machine classifier.

UNITED STATES NON-PROVISIONAL PATENT APPLICATION
FOR
SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB

Inventors:

Revanth Patil

Paulo Shakarian

Ashkan Aleali

Ericsson Marin

Status:

Micro-Entity

Priority:

U.S. Provisional Patent Application No. 62/668,878

Prepared by:

Polsinelli PC
One East Washington Street, Suite 1200
Phoenix, Arizona

Attorney's Docket No.:

055743-621457 (M18-191P)

SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This is a U.S. non-provisional patent application that claims benefit to U.S. provisional patent application serial number 62/668,878 filed on May 9, 2018, which is incorporated by reference in its entirety.

FIELD

[0002] The present disclosure generally relates to systems and methods for automatically assigning one or more labels or tags related to various discussion forum topics on the dark web or deep web.

BACKGROUND

[0003] The ability to traverse the Internet with complete anonymity provides online platforms for illegal activities such as credit card fraud, identity theft, leaks of sensitive information and sharing hacking information. One of the most prevalent cyber environments that emerged in the last decade and contributed to the achievement of those criminal tasks are dark web forums, since these forums include encryption technology to prevent monitoring and also provide protection from unauthorized users. Table 1 provides the details of thirteen broad categories of data seen in web based hidden services on the dark web, having discussion forums as the main supplier platform for the spread of criminal activities. For example, the deep web data can be classified into different domains such as hacking, whistle-blowing, financial-fraud, drugs, counterfeit, books, pornographic materials, etc.

Category	Percentage
Violence	0.3
Arms	0.8
Social	1.2
Hacking	1.8

Illegitimate pornography	2.3
Nexus	2.3
Extremism	2.7
Unknown	3.0
Other illicit	3.8
Finance	6.3
Drugs	8.1
Other	19.6
None	47.7

Table 1: Web Based Hidden Services in February 2016

[0004] Considering the enormity of the data in those environments, there is an impending need to go beyond this broad categorization, thereby providing to security researchers a more granular, structural and interdependent classification of the available information.

[0005] The dark web forms a small part of the deep web that is not indexed by web search engines. Dark web websites provide underground communication and are accessible by special software like Tor (The Onion Router) and I2p Invisible Internet Project. Tor software uses “onion routing” protocol to provide anonymity for both the service user and the service provider. Onion routing is a technique where messages are repeatedly encrypted and then sent through several network nodes, called onion routers. Like someone peeling an onion, each onion router removes a layer of encryption to uncover routing instructions, and sends the message to the next router where the process is repeated. This technique prevents intermediary nodes from knowing the origin, destination, and contents of the message.

[0006] Current technologies use learning models and techniques that do not address the issues of labeled data scarcity, nor do they address imbalanced data classes in the training set. Training sets are labeled by hand, which is a time-consuming and typically un-scalable process. A more thorough and versatile learning method is desired.

[0007] It is with these observations in mind, among others, that various aspects of the present disclosure were conceived and developed.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] FIG. **1** is a graphical representation showing frequency tags in a training set, according to aspects of the present disclosure;

[0009] FIG. **2** is a simplified illustration showing a simple Word2vec CBOW Model in the Only One Word in the Context, according to aspects of the present disclosure;

[0010] FIG. **3** is a simplified illustration showing a Word2vec CBOW framework for Learning Word Vectors, according to aspects of the present disclosure;

[0011] FIG. **4** is a simplified illustration showing a Doc2vec Pv-dm framework for Learning Vectors, according to aspects of the present disclosure;

[0012] FIGS **5A-5C** are simplified illustrations showing a sample of a tag-hierarchical structure, according to aspects of the present disclosure;

[0013] FIG. **6** is a simplified block diagram of a system overview, according to aspects of the present disclosure;

[0014] FIG. **7** is a graphical representation of a tag-wise precision scores for each experiment, according to aspects of the present disclosure;

[0015] FIG. **8** is a graphical representation of a tag-wise recall scores for each experiment, according to aspects of the present disclosure;

[0016] FIG. **9** is a graphical representation of a tag-wise F1 scores for each experiment, according to aspects of the present disclosure;

[0017] FIG. **10** is a graphical representation of a combination of adding parent tags and removing child tags with threshold, according to aspects of the present disclosure;

[0018] FIG. **11** is a graphical representation of a change in precision and recall scores for each tag vs. frequency of each tag in a training set for a Smote technique, according to aspects of the present disclosure;

[0019] FIG. **12** is a graphical representation of a change in precision and recall scores for each tag vs. frequency of each tag in a training set for Es (elastic search) based on semi-supervised technique, according to aspects of the present disclosure; and

[0020] FIG. **13** is a simplified block diagram of an exemplary computer system for effectuating the functions of the framework for security strength and performance analysis, according to aspects of the present disclosure.

[0021] Corresponding reference characters indicate corresponding elements among the view of the drawings. The headings used in the figures do not limit the scope of the claims.

DETAILED DESCRIPTION

[0022] The present disclosure is related to dark web-deep web (d2web) forums that are anonymously hosted and are identified by the domain '.onion'. Forums are online discussion sites where like-minded individuals can hold conversations. These anonymous hosting forums discuss various cyber-security related topics such as worms, botnet, zero-days, hacking tools, backdoor services and etc.

[0023] Aspects of the present disclosure include an inventive computer-implemented system (hereinafter "system" or "present system") that involves automatically assigning one or more labels (tags), in a hierarchical structure, to discussion topics seen in d2web forums. In some embodiments, the present system is formed by four main modules. In one method, the present system crawls various d2web sites and extracts important information from html pages, storing it in a database. Specialized crawlers automatically connect and authenticate the d2web sites through Tor. The important information is then parsed from these sites, such as discussion topics, user related information, discussion posts, etc., and the information stored on a database as well as on an elastic search data store. Data preprocessing and feature extraction is then performed over the forum discussion topics. This may be accomplished by eliminating all non-English words and irrelevant characters, conducting tokenization and stop word removal, stemming, and lemmatization. The present system then extracts features from every topic on the forum using Doc2vec. Finally, in the fourth module, different machine classifiers are used to assign multiple tags to each topic, so that their performances can be compared. A further description of the present system for automatically assigning one or more labels to d2web discussion topics is described in greater detail below.

Problem Statement

[0024] Classification of d2web forum discussions is a multiclass, multi-label classification problem, where the goal is to automatically assign one or more labels to each discussion topics. Formally, let X be a set of forum topics, $y = \{0, 1\}^k$, let k be possible tags and D be an unknown distribution on the product space $X \times Y$. Each element (x,y) in this space is composed of an instance x and a label vector y, which is a vector of indicators $y = [y_1, \ldots, y_k]$ that specifies classes associated with x. A classifier is a function h: $X \rightarrow Y$, that maps instance x to a label vector y =

h(x). For example, the sample forum topic “warning enigma badlock - upcoming smb/cifs and samba vulnerability” belongs to three classes – “vuln”, “smb” and “file-share”.

[0025] The present system leverages data crawled from a dataset of 283 cyber security related online forums. From the collected html pages or documents of the dataset, the present system parses predetermined important fields from the data such as topic title, posts, user name, title posted date, user ratings, number of replies, etc. The dataset includes discussions related to 486,996 different topics with 4,188,345 posts and 748,698 users participating in those topics.

[0026] The ground truth to build a machine learning model may be hand labeled by field experts. The present system has 2,046 labeled topics as a training set and they belong to 226 unique tags. FIG. **1** shows the tag frequency in the training set. Some tags have more than 400 topics in the training set while some have just one topic. Table 2 provides the sample ground truth topics and the tags associated.

Feature Extraction

[0027] Textual feature extraction was performed on topic title using a Doc2vec vectorization technique. Doc2vec is an extension of Word2vec that relies on the idea that words which appear in the similar context have the same semantic meaning. Word2vec is computationally-efficient predictive model that uses shallow two layer neural network for learning word embeddings. A Distributed Memory Model of Paragraph Vectors (PV-DM) Doc2vec architecture was used which is based on Continuous Bag Of Words (CBOW) Word2vec architecture. PV-DM architecture is faster to train and has better accuracy for frequent words.

Topic titles	Tags
enigma badlock - upcoming smb cifs and samba vulnerability.	vuln, smb, file-share
u nessus vulnerability scan	vuln, pentest, network
warning enigma badlock - upcoming smb/cifs and samba vulnerability.	vuln, smb, file-share
question about school-sponsored wifi?	wifi, inquiry

technical question about wifi cards	wifi, inquiry
(b) paypal unchecked accounts (s) btc account	financial-fraud, darkweb product, paypal, payment-services, market
there are phishing attempts reported!	phishing
[u'[payload] sharing some payloads!']	malware, payload, hack

Table 2: Sample Ground Truth

[0028] **Word2vec CBOW.** In this architecture, the model implemented by the system predicts the current word from a sliding window of context words. FIG. **2** shows a CBOW architecture with just one word considered per context. Consider the settings with V as vocabulary size and N as hidden layer size. The input is a one-hot encoded vector (i.e. for a given input context word, only one of the V units, {x1, x2,...., xv}, is 1). Weights between the input and the hidden layer are represented by W which is of dimensions $V \times N$. The activation function for the hidden layer is linear. Formally, row i of W and for a given context word w,

[0029] Equation: 1

$$h = \mathbf{W}^{\mathrm{T}} \times x := \mathbf{v}_w^{\mathrm{T}}$$

v_w is the the vector representation of the input word w. From the hidden layer to the output layer, there is different weight matrix W' which is of dimension$N \times V$. A score u_j is computed for each word in the vocabulary,

$$u_j = \mathbf{v}_{w_j}^{\prime\mathrm{T}}\mathbf{h}$$

where $\mathbf{v}_{w_j}^{\prime\mathrm{T}}$ is the j-th column of the matrix $\mathbf{W}'$. A log-linear softmax function is applied to obtain posterior distribution of words. Formally, y_j is the output of jth unit in the output layer.

$$y_j = \frac{exp(u_j)}{\sum_{j'=1}^{V} \exp(u_{j_l})}$$

[0030] The weights ($\mathbf{W}$ and$\mathbf{W}'$) are updated using a logarithmic loss function with a back propagation method. After updating the weights, $\mathbf{v}_w$ is used for word w from Equation 1 as the 'input vector' word embedding.

[0031] **Doc2vec PV-DM**. A paragraph vector is a supervised framework that may be implemented by the system that learns continuous distributed vector representations for documents instead of words. The text in the documents can be of variable length. As shown in FIG. **4**, every paragraph is mapped to a unique vector, represented by a column matrix *D* and every word is also mapped to a unique vector, represented by a column in matrix *W*. The paragraph vector and word vectors are averaged or concatenated to predict the next word in a context. The only difference between this Doc2vec PV-DM model compared to the Word2vec framework is the addition of a paragraph vector along with the word vectors.

Tags and the Hierarchy among Tags

[0032] In the ground truth, each topic is associated with multiple tags. It was observed that there is a natural hierarchical structure in the set of tags. There are group tags (parent tags) that can be seen as the "broader term" for its set of specific tags (child tags). Nesting specific tags under group tags creates a hierarchy of tags. FIGS. **5A-5C** show an example of the hierarchical structure in the training dataset. In the illustrated case (FIG. **5B**), external-world tag is the more generic parent tag for social-realm and health tag. Similarly social-realm and health tags are child tags to external-world tag.

[0033] Formally, a hierarchy constraint is given by a set $\mathcal{H} = \{H_1, H_2, \ldots . H_n\}$ where H_i is a tree. For all the tags *T*, a set of tags *S* ($S \subseteq T$) are said to be consistent if every tag t ($t \subset S$) satisfies either of the two conditions:

1. $p_t = t\ i.e\ t$ is a root node
2. p_t is consistent with respect to $\mathcal{H}$

where p_t tag is parent of tag t.

Baseline Approach

[0034] For the baseline approach, Support Vector Machine, decision tree, and random forest classifiers are used to perform multi-class and multi-label classification.

[0035] A Support Vector Machine (SVM) is a supervised learning method that finds a hyper-plane that may differentiate two classes in a multi-

dimensional feature space. In addition to performing linear classification, SVM can efficiently perform a non-linear classification using kernel function, implicitly mapping their inputs into high-dimensional feature spaces. The SVM radial basis function (rbf) kernel is used with "one-vs-the-rest" strategy to perform multi-class multi-label classification.

[0036] Decision trees are a non-parametric supervised learning method used for classification and regression. These decision trees learn the simple decision rules from the data features to predict the target label. The decision tree method is a recursive partitioning algorithm that is widely used for classification problems. This algorithm aims to maximize the information gain at every step.

[0037] Random Forest classifier is an ensemble algorithm that creates different decision trees by randomly selecting a subset of a training set and a subset of features. The number of features that are searched at each split point is specified as the parameter. To reduce memory consumption, the total number of trees and size of the trees is controlled. Random Forest then aggregates the votes from different decision trees to decide the final class of the test object.

Overview of Problem Specific Challenges

[0038] **Ensuring Consistency in Results Based on Hierarchy.** In the ground truth, it is observed that the tags had natural hierarchy. As the original ground truth is hand labeled, it cannot expect all of the ground truth entries to respect the tag hierarchy due to human error. Therefore, an automation script was developed that establishes the hierarchy in topic titles in the ground truth. Now, with hierarchy established in the tags, the classifier provided detailed tag prediction for all the documents.

[0039] Ideally, the classifier should predict the hierarchical representation of tags, for example, for a predicted child tag every parent tag should also be predicted. However it is observed that for some tags returned by the classifier function, the corresponding parent tags is not returned. This happens when either the tag is wrongly predicted (false positive) or when some of the correct tags (false negatives) are not predicted.

[0040] **Class Imbalance:** The class imbalance problem typically occurs when, in a classification problem, there are many more instances of some classes than the others. In such cases, classifiers tend to be overwhelmed by the majority

classes and ignore the minority classes. These minority classes usually have lesser precision and recall scores in the baseline approach.

[0041] Semi-supervised learning techniques are usually used to over-sample and under-sample the minority or majority class for adjusting the class distribution of a dataset.

Leveraging Tag Hierarchy

[0042] Three approaches are used (adding parent tag, removing child tag, and a combination of adding parent and removing child tag) to ensure that tag hierarchy is maintained in the prediction list.

[0043] **Adding Parent Tag**. If a child tag is predicted by the classifier and if the corresponding parent tags in the hierarchy are not predicted, then all the parent tags are added to the prediction list. In FIG. **5B**, with the tag hierarchy – "external-world", "social-realm", "health" – if a "health" tag is predicted without "external-world" and "social-realm", then "external-world" and "social-realm" tags are added to the prediction list.

[0044] **Removing Child Tag**. In this approach, all predicted child tags are removed that do not have all the hierarchical parent tags. In FIG. **5B**, with the tag hierarchy – "external-world", "social-realm", "health" - if "external world" and "health" tags are predicted without "social-realm", then the "health" tag is removed from the prediction list, keeping only "external-world".

Algorithm 1 Combination of Adding Parent Tags and Removing Child Tags (**T**)

Input: Prediction list of tags (T) for each document.

Output: Prediction list of tags that respect the tag hierarchy for each document.

1: $\beta \leftarrow$ remove child threshold value

2: $\alpha \leftarrow$ add parent threshold value

3: **for** t:= 1 to T **do**

4: *p(t)* $\leftarrow$ prediction probability of tag t.

5: **if** p(t) < β **then** *remove all child tags of tag t from prediction list*

6: **for** t:= 1 to T **do**

7: p(t) ← prediction probability of tag t.

8: **if** p(t) > α **then** *add all parent tags of tag t from prediction list*

return updated prediction list

[0045] Combination of Adding Parent Tag and Removing Child Tag

In this approach, the present system regulates the way parent tags are added and child tags are removed, based on the probability of the predicted tag. The following threshold values are used - 'add parent threshold' (α) and 'remove child threshold' (β) - to decide whether to add parent tag or remove child tag. The procedure for this approach is described in Algorithm 1.

Dealing with Class Imbalance

[0046] The discussions in the dark forums are based on various topics which are unequally distributed. Therefore, class imbalance is observed in the ground truth. In the dataset, the ratio of the minority to the majority classes is usually around 1:400, which is quite drastic. To deal with the class imbalance problem, semi-supervised learning methods are utilized. Semi-supervised learning is a class of supervised learning techniques that also make use of additional unlabeled data to better capture the shape of the underlying data distribution and generalize new samples more effectively.

Synthetic Over-Sampling Technique

[0047] Synthetic Minority Over-Sampling Technique (SMOTE) is used as a baseline approach to solve the class imbalance problem. Oversampling is to correct for a bias in the original dataset. In SMOTE, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen.

[0048] For each minority class sample synthetic examples are introduced along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. The present system currently uses five nearest neighbors. Synthetic samples are generated in the following way:

1) Take the difference between the feature vector under consideration and its nearest neighbor.

2) Multiply this difference by a random number between zero and one.

3) Add it to the feature vector under consideration.

In this manner, a random point is synthetically introduced along the line segment between two specific features.

Leveraging Document Similarity through Elastic Search

[0049] One way to solve the class imbalance problem is by increasing the ground truth size, but that is difficult as a human-labeled data-set is used. Therefore, a semi-supervised method is used to increase the sample size of the minority classes. Semi-supervised learning methods also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data. For every sample in the minority class, we find the top similar documents from the database using the elastic search similarity score and add them to the training data.

[0050] Elastic search is a high-performance, full-featured text search engine library. Its ranking function is applied to determine how relevant a retrieved document is to a given query. The ranking function is based on a combination of Vector Space Model and a Boolean model of Information Retrieval. The main idea behind this approach is the more times a query term appears in a document relative to the number of times the term appears in the whole collection, the more relevant that document will be to the query. The elastic search uses BM25 as the default ranking function to provide the relevance score to a given search query. BM25 is not a single function, rather, a family of TF-IDF like retrieval functions. One of the most commonly used scoring functions is described below. Given a query Q, containing keywords q1,..,qn, the BM25 score of a document D is

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i).\frac{f(q_i,D).(k_1+1)}{f(q_i,\ D + k_1.(1-b+b.\frac{|D|}{avgdl})}$$

[0051] where $f(q_i, D)$ is q_i's term frequency in document D, |*D*| is the length of the document D in words and avgdl is the average document length in the text collection from which documents are drawn. k_1 and b are free parameters,

usually chosen, in absence of an advanced optimization, as $k_1 \in [1.2, 2.0]$ and b = 0.75. IDF(q_i) is the Inverse document frequency weight of the query term q_i. IDF(q_i) is usually computed as:

$$IDF(q_i) = log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

where N is the total number of documents in the collection, and n(q_i) is the number of documents containing q_i. The semi-supervised relevance score based is described by algorithm 2.

Algorithm 2 Semi-Supervised Learning with Document Similarity (**T**, **Tr** (T_i, T_p), **F**)

Input: Tags in training set T, Training set Tr with tuple (document titles T_i, tag predictions T_p, Frequency of tags in training set F.

Output: Increased minority class samples from untrained dataset, Ims.

1: T_{min} ← array of minority classes
2: **for** t in T **do**
3: if *F*(*t*) < 150 **then**
4: T_{min} ← *add tag t*
5: **for** ti in Ti **do**
6: *tp* ← $T_p(ti)$
7: **if** *tp* in T_{min} **then**
8: sd ← use relevance score and find top similar documents from untrained data
9: *Ims* ← add tuple(sd, tp)
10: *Ims* ← add Tr
return *Ims*

EXPERIMENTS

[0052] FIG. **6** shows the overview of the present system **100**. As shown, the present system **100** consists of four main modules - data collection **102**, data-preprocessing **104**, feature extraction **106**, and classification **108**. Spider programs are used to crawl dark web forums and human analysts filter these forums that are relevant to cyber security domain. On these forum sites, Tor-based specialized crawlers are used to traverse the website and retrieve Internet documents and related information, such as HTML pages. Each site has a parser program that extracts important information, such as fields including a topic title, topic content, post content, topic author, post author, author reputation, posted date for each title and posts, etc., from HTML pages, all of which is stored on a database. In some embodiments, expert analysts use a fraction of this data to prepare ground truth by labeling tags to topic titles. In some embodiments, this ground truth may form the basis for the next 3 modules to develop the machine learning model.

[0053] In addition, FIG. **6** shows the data preprocessing **104** steps - eliminating duplicates in the ground truth, eliminating non-English content in both the ground truth and the test data, tokenizing the data, removing stop words, and applying stemming and lemmatization processes on the data.

[0054] **Tokenization**. Given a sequence of characters, tokenization is the task of chopping it up into pieces, called tokens, at the same time throwing away certain characters, such as punctuation and non-asci characters. In the meanwhile, the useful characters are kept, such as currency symbols, question mark etc. that add semantic meaning to the features.

[0055] **Stop Word Removal.** Extremely common words that appear in most of the documents are of little value in feature extraction. Therefore, these common words are removed from the topic titles.

[0056] **Stemming and Lemmatization**. Stemming is the process for reducing inflected or derived words to their stem, base or root form. Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization and stemming are closely related. The small difference is that lemmatization uses the context, whereas stemming operates on a single word. For example, stemming process reduces the derived words – “caresses”, “ponies” and “cats” to “caress”, “poni”, and “cat” respectively. On the

other hand, lemmatization reduces the derived word “saw” to “see” or “saw” depending on whether the use of token was a verb or a noun.

Description of Software

[0057] Table 3 provides a summary of a software application **150** that may be developed and implemented to perform aspects of the system **100** described herein, and version description for each of the possible modules used in the application **150**. In one embodiment, for data-preprocessing, a module textblob with version‘0.12.0’ was used for part-of-speech tagging and natural language toolkit (nltk) library with version ‘3.2.4’ for tokenization, stemming and lemmatization. For vectorization, scikit-learn and sklearn version ‘0.19.0’ were used for Tf-IDF feature extraction and gensim module version ‘2.3.0’ for doc2vec vectorization. For supervised classification, svm was used, decision tree and random forest classifiers built in sklearn ‘0.19.0’. For SMOTE technique, an imblearn module version ‘0.3.0’ was used. All the above modules of the application **150** may be developed using python ‘2.7.13’ programming language, but the present disclosure is not limited in this regard. To find similar documents using ES, the elastic search version ‘5.3.2’ with backend lucene search engine version ‘6.4.2’ was used. All the systems were implemented in Linux based Ubuntu 16.04 LTS operating system.

Module	**Software**
Data preprocessing	textblob ‘0.12.0’
Tokenization, Stemming and Lemmatization	Natural language toolkit(nltk) ‘3.2.4’
Doc2vec	Gensim ‘2.3.0’
SVM, the decision tree and random forest classifications	scikit-learn ‘0.19.0’
SMOTE	imblearn ‘0.3.0’

Table 3: Software Description of the Modules (for application **150** shown in FIG. **13**).

Description of Metrics

[0058] The performance of the classifier models is evaluated based on four metrics - precision, recall, F1 score, and percentage of documents titles with at

least one correct prediction tag. Precision, recall and F1 scores are calculated for each of the tags and for all the tags put together. Table 4 lists the formal definitions of precision, recall and F1 scores. Precision score is defined as the fraction of correctly predicted document titles from all the predicted document titles. Recall score is defined as the fraction of correctly predicted document titles from the total number of document titles. F1 measure is the harmonic mean of precision and recall score. The precision, recall, and F1 scores were calculated for individual tags and for the cumulative of all the tags.

Metric	Formula
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1	$2 * \frac{precision * recall}{precision + recall}$

Table 4: EVALUATION METRICS. True Positives (TOP), False Positives (FP), False Negatives (FN).

[0059] The results of different experiments; baseline classifiers, classifiers with tag hierarchy considered, and classifiers with semi supervised learning technique are described below.

Results of Baseline Classifier

[0060] In all the experiments, a training set with topic titles to perform 10-fold cross validation is used to validate the models. With 10-fold cross validation, the ground truth data is randomly partitioned into 10 equal subsample buckets. Out of 10 buckets, one bucket is used for testing the model and the rest of the ground truth is used for training the model. Each of the 10 bucket subsamples are used for testing the model in cross validation. The aggregate of the accuracy scores of k scores is used as the final accuracy score. As the ground truth data contains 2,046 titles, each of the 10 buckets consists of 204 samples for testing and 1,842 for training the model.

Experiment	Precision %	Recall %	F1 score %	Percentage of documents with at least one correct prediction
Baseline Decision trees	58.3	6.41	11.54	24
Baseline SVM	67.2	8.51	15.1	27.9
Baseline Random Forest	71.1	11.3	19.7	31.37
Add parent	71.9	20.2	31.5	46.07
Remove Child	74	11.4	19.47	43.6
SMOTE	40.3	30.1	34.4	61.76
Semi-supervised learning based on Elastic Search relevance score	78.4	21.2	33.1	54.9

Table 5: Experimental Results

[0061] Decision Tree, SVM, and Random Forest were used to perform multi-class and multi-label classification. For the decision tree classifier, pruning to avoid overfitting was used, by setting the minimum number of samples required at a leaf node to 5%. For SVM kernel, a 'one-vs-the-rest' strategy was used with radial basis function (rbf) kernel. For the random forest, the number of estimators of the tree parameter was tuned and found optimal results when number of estimators is 200. In the above mentioned setting, the Random Forest performed best with the 71.1% precision score and 11.3% recall score. Therefore, Random Forest was used as the default classifier for all the experiments.

Change in Precision and Recall Scores for Each Tag Vs Frequency of Each Tag in Training Set for ES Based Semi-supervised Technique

[0062] Three different approaches were used - adding Parent tag, removing child tag, combination of adding parent and removing child tag - to preserve the tag hierarchy in all tag predictions.

Adding Parent Tags

[0063] In every iteration of the K-fold cross validation, if a child tag is predicted by the classifier and if the corresponding parent tags in the hierarchy are

not predicted, then all the parent tags are added to the prediction list. Adding parent tags decreases false negatives and increases true positives. Hence, the recall score would increase, but if the child tag is wrongly predicted then the error is propagated and increases the false positives, thereby decreasing the precision score. This observation reveals that there is a slight decrease in the precision score, but the recall score improved significantly from 11.3% to 20.2%.

Removing Child Tags

[0064] In every iteration of K-fold cross validation, all predicted child tags that did not have all the corresponding hierarchical parent tags are removed. This experiment reduces the false positives, thereby increasing the precision. The observation from this experiment revealed that the precision increased from 71.1% to 74%, and the recall score did not change from the baseline.

Combination of Adding Parent Tags and Removing Child Tags (CAR)

[0065] In this approach, the desire was to regulate the way parent tags were added and child tags removed based on the probability of the predicted tag. The "add parent threshold" (α) or the "remove child threshold values" (β) are used in order to decide on whether to add or remove a particular tag, respectively.

[0066] Removing child tags reduces false positives, increasing the precision score, while adding parent tags reduces false negatives, increasing the recall score. The above experiment was repeated for different "remove child threshold" (β) and "add parent threshold" (α) values. As shown in FIG. **10**, with an increase in "add parent threshold" (α) value, the recall score increased and precision score decreased. Conversely, with an increase in "remove child threshold" value (β), the precision score increased and recall score decreased. The aim was to find an optimum increase in F1 score without drastically affecting the precision score by tuning the (α) and (β) values. The best F1 score was achieved when β = 0.5 and α = 0.9. The corresponding F1 score is 31.9%, precision score is 73.43% and recall score is 20.4%.

Results for Experiments on Class Imbalance

[0067] Semi-supervised learning methods were used to solve class imbalance problem by making use of unlabeled data for training. For both SMOTE and elastic search, we use the ground truth with tag hierarchy and preserve the tag hierarchy in prediction list by using a 'combination of adding parent tags and removing child tags' method with α = 0.5 and β = 0.9.

SMOTE Semi-Supervised Technique

[0068] In SMOTE, synthetic samples are introduced along the line segments joining all of the k minority class nearest neighbors. For experimental data, the number of nearest neighbors to construct synthetic samples used is 5. The targeted minority classes will be over-sampled to achieve an equal number of samples across all classes. With the introduction of synthetic classes, a reduction in false negatives and an increase in false positives are observed. Therefore, recall for SMOTE increased from 11.3% to 30.1% and precision decreased from 71.1% to 40%. The overall F1 score increased from 19.7% to 34.4%.

[0069] As shown in FIG. **11**, SMOTE technique was most effective in increasing the recall score for the tags that have frequencies between 95 to 300 in the ground truth. For this range, the average increase in recall for a tag is 23.35%. However, this technique is not very useful for the tags that have a frequency fewer than 95 or for the tags that have a frequency higher than 300. For these ranges, the average increase in recall is just 8.33%. As shown in the Table 5, the percentage of documents with at least one correct prediction increased from 31.37% to 61.76% with SMOTE semi-supervised learning.

Elastic Search Based Semi-Supervised Technique

[0070] During experimental testing, the training samples were extended of the each minority class by using document similarity score.

[0071] Using the above procedure, the training samples of each of the minority classes is extended to have 500 document titles each. With this approach, precision increases from 71.1% to 78.4%, recall increases from 11.3% to 21.1%, and F1 increases from 19.7% to 33.1%. Compared to SMOTE, precision increases from 40% to 78.4% but there is a dip in recall from 30.1% to 21.1%. When augmenting synthetic samples with the training set using SMOTE technique, it was observed that

false negatives decrease and false positives increase with the introduction of synthetic samples. However with Elastic Search-based technique, the false negatives decrease and true positives increase.

[0072] As shown in FIG. **12**, precision and recall increase for most of the tags. This technique is most effective for the tags that had frequencies between 86 to 318. For this range the average increase in precision is 52% and average increase in recall is 12%. For the tags having frequencies below 86, the increase in precision and recall is just 2.86% and 5.6% respectively. As shown in Table 5, the percentage of documents with at least one correct prediction increases from 31.37% to 54.9% with this approach.

Computer-implemented System

[0073] FIG. **13** illustrates an example of a suitable computing and networking environment (computer system **200**) which may be used to implement various aspects of the present disclosure, such as the application **150**. Example embodiments described herein may be implemented at least in part in electronic circuitry; in computer hardware executing firmware and/or software instructions; and/or in combinations thereof. Example embodiments also may be implemented using a computer program product (e.g., a computer program tangibly or non-transitorily embodied in a machine-readable medium and including instructions for execution by, or to control the operation of, a data processing apparatus, such as, for example, one or more programmable processors or computers). A computer program may be written in any form of programming language, including compiled or interpreted languages, and may be deployed in any form, including as a stand-alone program or as a subroutine or other unit suitable for use in a computing environment. Also, a computer program can be deployed to be executed on one computer, or to be executed on multiple computers at one site or distributed across multiple sites and interconnected by a communication network.

[0074] Certain embodiments are described herein as including one or more modules. Such modules are hardware-implemented, and thus include at least one tangible unit capable of performing certain operations and may be configured or arranged in a certain manner. For example, a hardware-implemented module may comprise dedicated circuitry that is permanently configured (e.g., as a special-purpose processor, such as a field-programmable gate array (FPGA) or an

application-specific integrated circuit (ASIC)) to perform certain operations. A hardware-implemented module may also comprise programmable circuitry (e.g., as encompassed within a general-purpose processor or other programmable processor) that is temporarily configured by software or firmware to perform certain operations. In some example embodiments, one or more computer systems (e.g., a standalone system, a client and/or server computer system, or a peer-to-peer computer system) or one or more processors may be configured by software (e.g., an application or application portion) as a hardware-implemented module that operates to perform certain operations as described herein.

[0075] Accordingly, the term "hardware-implemented module" encompasses a tangible entity, be that an entity that is physically constructed, permanently configured (e.g., hardwired), or temporarily configured (e.g., programmed) to operate in a certain manner and/or to perform certain operations described herein. Considering embodiments in which hardware-implemented modules are temporarily configured (e.g., programmed), each of the hardware-implemented modules need not be configured or instantiated at any one instance in time. For example, where the hardware-implemented modules comprise a general-purpose processor configured using software, the general-purpose processor may be configured as respective different hardware-implemented modules at different times. Software, in the form of the application **150** or otherwise, may include a hardware-implemented module and may accordingly configure a processor **202**, for example, to constitute a particular hardware-implemented module at one instance of time and to constitute a different hardware-implemented module at a different instance of time.

[0076] Hardware-implemented modules may provide information to, and/or receive information from, other hardware-implemented modules. Accordingly, the described hardware-implemented modules may be regarded as being communicatively coupled. Where multiple of such hardware-implemented modules exist contemporaneously, communications may be achieved through signal transmission (e.g., over appropriate circuits and buses) that connect the hardware-implemented modules. In embodiments in which multiple hardware-implemented modules are configured or instantiated at different times, communications between such hardware-implemented modules may be achieved, for example, through the storage and retrieval of information in memory structures to which the multiple

hardware-implemented modules have access. For example, one hardware-implemented module may perform an operation, and may store the output of that operation in a memory device to which it is communicatively coupled. A further hardware-implemented module may then, at a later time, access the memory device to retrieve and process the stored output. Hardware-implemented modules may also initiate communications with input or output devices.

[0077] As illustrated, the computing and networking environment **200** may be a general purpose computing device **200**, although it is contemplated that the networking environment **200** may include other computing systems, such as personal computers, server computers, hand-held or laptop devices, tablet devices, multiprocessor systems, microprocessor-based systems, set top boxes, programmable consumer electronic devices, network PCs, minicomputers, mainframe computers, digital signal processors, state machines, logic circuitries, distributed computing environments that include any of the above computing systems or devices, and the like.

[0078] Components of the general purpose computing device **200** may include various hardware components, such as a processing unit **202**, a main memory **204** (e.g., a system memory), and a system bus **201** that couples various system components of the general purpose computing device **200** to the processing unit **202**. The system bus **201** may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. For example, such architectures may include Industry Standard Architecture (ISA) bus, Micro Channel Architecture (MCA) bus, Enhanced ISA (EISA) bus, Video Electronics Standards Association (VESA) local bus, and Peripheral Component Interconnect (PCI) bus also known as Mezzanine bus.

[0079] The general purpose computing device **200** may further include a variety of computer-readable media **207** that includes removable/non-removable media and volatile/nonvolatile media, but excludes transitory propagated signals. Computer-readable media **207** may also include computer storage media and communication media. Computer storage media includes removable/non-removable media and volatile/nonvolatile media implemented in any method or technology for storage of information, such as computer-readable instructions, data structures, program modules or other data, such as RAM, ROM, EPSOM, flash memory or other

memory technology, CD-ROM, digital versatile disks (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium that may be used to store the desired information/data and which may be accessed by the general purpose computing device **200**. Communication media includes computer-readable instructions, data structures, program modules, or other data in a modulated data signal such as a carrier wave or other transport mechanism and includes any information delivery media. The term "modulated data signal" means a signal that has one or more of its characteristics set or changed in such a manner as to encode information in the signal. For example, communication media may include wired media such as a wired network or direct-wired connection and wireless media such as acoustic, RF, infrared, and/or other wireless media, or some combination thereof. Computer-readable media may be embodied as a computer program product, such as software stored on computer storage media.

[0080] The main memory **204** includes computer storage media in the form of volatile/nonvolatile memory such as read only memory (ROM) and random access memory (RAM). A basic input/output system (BIOS), containing the basic routines that help to transfer information between elements within the general purpose computing device **200** (e.g., during start-up) is typically stored in ROM. RAM typically contains data and/or program modules that are immediately accessible to and/or presently being operated on by processing unit **202**. For example, in one embodiment, data storage **206** holds an operating system, application programs, and other program modules and program data.

[0081] Data storage **206** may also include other removable/non-removable, volatile/nonvolatile computer storage media. For example, data storage **206** may be: a hard disk drive that reads from or writes to non-removable, nonvolatile magnetic media; a magnetic disk drive that reads from or writes to a removable, nonvolatile magnetic disk; and/or an optical disk drive that reads from or writes to a removable, nonvolatile optical disk such as a CD-ROM or other optical media. Other removable/non-removable, volatile/nonvolatile computer storage media may include magnetic tape cassettes, flash memory cards, digital versatile disks, digital video tape, solid state RAM, solid state ROM, and the like. The drives and their associated computer storage media provide storage of computer-readable instructions, data

structures, program modules and other data for the general purpose computing device **200**.

[0082] A user may enter commands and information through a user interface **240** or other input devices **245** such as a tablet, electronic digitizer, a microphone, keyboard, and/or pointing device, commonly referred to as mouse, trackball, or touch pad. Other input devices **245** may include a joystick, game pad, satellite dish, scanner, or the like. Additionally, voice inputs, gesture inputs (e.g., via hands or fingers), or other natural user interfaces may also be used with the appropriate input devices, such as a microphone, camera, tablet, touch pad, glove, or other sensor. These and other input devices **245** are often connected to the processing unit **202** through a user interface **240** that is coupled to the system bus **201**, but may be connected by other interface and bus structures, such as a parallel port, game port or a universal serial bus (USB). A monitor **260** or other type of display device is also connected to the system bus **201** via user interface **240**, such as a video interface. The monitor **260** may also be integrated with a touch-screen panel or the like.

[0083] The general purpose computing device **200** may operate in a networked or cloud-computing environment using logical connections of a network Interface **203** to one or more remote devices, such as a remote computer. The remote computer may be a personal computer, a server, a router, a network PC, a peer device or other common network node, and typically includes many or all of the elements described above relative to the general purpose computing device **200**. The logical connection may include one or more local area networks (LAN) and one or more wide area networks (WAN), but may also include other networks. Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets and the Internet.

[0084] When used in a networked or cloud-computing environment, the general purpose computing device **200** may be connected to a public and/or private network through the network interface **203**. In such embodiments, a modem or other means for establishing communications over the network is connected to the system bus **201** via the network interface **203** or other appropriate mechanism. A wireless networking component including an interface and antenna may be coupled through a suitable device such as an access point or peer computer to a network. In a networked environment, program modules depicted relative to the general purpose

computing device **200**, or portions thereof, may be stored in the remote memory storage device.

[0085] It should be understood from the foregoing that, while particular embodiments have been illustrated and described, various modifications can be made thereto without departing from the spirit and scope of the invention as will be apparent to those skilled in the art. Such changes and modifications are within the scope and teachings of this invention as defined in the claims appended hereto.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
16/405,612	05/07/2019	Revanth Patil	055743-621457 (M18-191P)

CONFIRMATION NO. 4562

27148
POLSINELLI PC
900 WEST 48TH PLACE
SUITE 900
KANSAS CITY, MO 64112-1895

INFORMAL NOTICE

OC00000108278675

Date Mailed: 05/20/2019

INFORMATIONAL NOTICE TO APPLICANT

Applicant is notified that the above-identified application contains the deficiencies noted below. No period for reply is set forth in this notice for correction of these deficiencies. However, if a deficiency relates to the inventor's oath or declaration, the applicant must file an oath or declaration in compliance with 37 CFR 1.63, or a substitute statement in compliance with 37 CFR 1.64, executed by or with respect to each actual inventor no later than the expiration of the time period set in the "Notice of Allowability" to avoid abandonment. See 37 CFR 1.53(f).

The item(s) indicated below are also required and should be submitted with any reply to this notice to avoid further processing delays.

- A properly executed inventor's oath or declaration has not been received for the following inventor(s):
  Revanth Patil
  Paulo Shakarian
  Ashkan Aleali
  Ericsson Marin

Questions about the contents of this notice and the requirements it sets forth should be directed to the Office of Data Management, Application Assistance Unit, at **(571) 272-4000** or **(571) 272-4200** or **1-888-786-0101**.

/hmarzi/

IPR2025-01398
CrowdStrike EX1002 Page 61

PATENT APPLICATION FEE DETERMINATION RECORD

Substitute for Form PTO-875

Application or Docket Number: 16/405,612

APPLICATION AS FILED - PART I

FOR	(Column 1) NUMBER FILED	(Column 2) NUMBER EXTRA	SMALL ENTITY RATE($)	SMALL ENTITY FEE($)	OR	OTHER THAN SMALL ENTITY RATE($)	OTHER THAN SMALL ENTITY FEE($)
BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A	N/A			N/A	75
SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A	N/A			N/A	165
EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A	N/A			N/A	190
TOTAL CLAIMS (37 CFR 1.16(i))	20 minus 20 =	*			OR	x 25 =	0.00
INDEPENDENT CLAIMS (37 CFR 1.16(h))	3 minus 3 =	*				x 115 =	0.00
APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $310 ($155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).						0.00
MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))							0.00
* If the difference in column 1 is less than zero, enter "0" in column 2.			TOTAL			TOTAL	430

APPLICATION AS AMENDED - PART II

		(Column 1) CLAIMS REMAINING AFTER AMENDMENT		(Column 2) HIGHEST NUMBER PREVIOUSLY PAID FOR	(Column 3) PRESENT EXTRA	SMALL ENTITY RATE($)	SMALL ENTITY ADDITIONAL FEE($)	OR	OTHER THAN SMALL ENTITY RATE($)	OTHER THAN SMALL ENTITY ADDITIONAL FEE($)
AMENDMENT A	Total (37 CFR 1.16(i))	*	Minus	**	=	x =		OR	x =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x =		OR	x =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

		(Column 1) CLAIMS REMAINING AFTER AMENDMENT		(Column 2) HIGHEST NUMBER PREVIOUSLY PAID FOR	(Column 3) PRESENT EXTRA	RATE($)	ADDITIONAL FEE($)		RATE($)	ADDITIONAL FEE($)
AMENDMENT B	Total (37 CFR 1.16(i))	*	Minus	**	=	x =		OR	x =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x =		OR	x =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".
The "Highest Number Previously Paid For" (Total or Independent) is the highest found in the appropriate box in column 1.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING or 371(c) DATE	GRP ART UNIT	FIL FEE REC'D	ATTY.DOCKET.NO	TOT CLAIMS	IND CLAIMS
16/405,612	05/07/2019	2144	470	055743-621457 (M18-191P)	20	3

CONFIRMATION NO. 4562

FILING RECEIPT

27148
POLSINELLI PC
900 WEST 48TH PLACE
SUITE 900
KANSAS CITY, MO 64112-1895

OC000000108278674

Date Mailed: 05/20/2019

Receipt is acknowledged of this non-provisional utility patent application. The application will be taken up for examination in due course. Applicant will be notified as to the results of the examination. Any correspondence concerning the application must include the following identification information: the U.S. APPLICATION NUMBER, FILING DATE, NAME OF FIRST INVENTOR, and TITLE OF INVENTION. Fees transmitted by check or draft are subject to collection.

Please verify the accuracy of the data presented on this receipt. If an error is noted on this Filing Receipt, please submit a written request for a corrected Filing Receipt, including a properly marked-up ADS showing the changes with strike-through for deletions and underlining for additions. If you received a "Notice to File Missing Parts" or other Notice requiring a response for this application, please submit any request for correction to this Filing Receipt with your reply to the Notice. When the USPTO processes the reply to the Notice, the USPTO will generate another Filing Receipt incorporating the requested corrections provided that the request is grantable.

Inventor(s)

Revanth Patil, Tempe, AZ;
Paulo Shakarian, Tempe, AZ;
Ashkan Aleali, Tempe, AZ;
Ericsson Marin, Tempe, AZ;

Applicant(s)

Revanth Patil, Tempe, AZ;
Paulo Shakarian, Tempe, AZ;
Ashkan Aleali, Tempe, AZ;
Ericsson Marin, Tempe, AZ;

Assignment For Published Patent Application

Arizona Board of Regents on Behalf of Arizona State University, Tempe, AZ

Power of Attorney: None

Domestic Priority data as claimed by applicant

This appln claims benefit of 62/668,878 05/09/2018

Foreign Applications for which priority is claimed (You may be eligible to benefit from the **Patent Prosecution Highway** program at the USPTO. Please see http://www.uspto.gov for more information.) - None.
Foreign application information must be provided in an Application Data Sheet in order to constitute a claim to foreign priority. See 37 CFR 1.55 and 1.76.

Permission to Access Application via Priority Document Exchange: Yes

Permission to Access Search Results: Yes

Applicant may provide or rescind an authorization for access using Form PTO/SB/39 or Form PTO/SB/69 as appropriate.

If Required, Foreign Filing License Granted: 05/17/2019

The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is **US 16/405,612**

Projected Publication Date: 11/14/2019

Non-Publication Request: No

Early Publication Request: No
**** MICRO ENTITY ****
Title

SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB

Preliminary Class

715

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications: No

PROTECTING YOUR INVENTION OUTSIDE THE UNITED STATES

Since the rights granted by a U.S. patent extend only throughout the territory of the United States and have no effect in a foreign country, an inventor who wishes patent protection in another country must apply for a patent in a specific country or in regional patent offices. Applicants may wish to consider the filing of an international application under the Patent Cooperation Treaty (PCT). An international (PCT) application generally has the same effect as a regular national patent application in each PCT-member country. The PCT process **simplifies** the filing of patent applications on the same invention in member countries, but **does not result** in a grant of "an international patent" and does not eliminate the need of applicants to file additional documents and fees in countries where patent protection is desired.

Almost every country has its own patent law, and a person desiring a patent in a particular country must make an application for patent in that country in accordance with its particular laws. Since the laws of many countries differ in various respects from the patent law of the United States, applicants are advised to seek guidance from specific foreign countries to ensure that patent rights are not lost prematurely.

Applicants also are advised that in the case of inventions made in the United States, the Director of the USPTO must issue a license before applicants can apply for a patent in a foreign country. The filing of a U.S. patent application serves as a request for a foreign filing license. The application's filing receipt contains further information and guidance as to the status of applicant's license for foreign filing.

Applicants may wish to consult the USPTO booklet, "General Information Concerning Patents" (specifically, the section entitled "Treaties and Foreign Patents") for more information on timeframes and deadlines for filing foreign

IPR2025-01398
CrowdStrike EX1002 Page 64

patent applications. The guide is available either by contacting the USPTO Contact Center at 800-786-9199, or it can be viewed on the USPTO website at http://www.uspto.gov/web/offices/pac/doc/general/index.html.

For information on preventing theft of your intellectual property (patents, trademarks and copyrights), you may wish to consult the U.S. Government website, http://www.stopfakes.gov. Part of a Department of Commerce initiative, this website includes self-help "toolkits" giving innovators guidance on how to protect intellectual property in specific countries such as China, Korea and Mexico. For questions regarding patent enforcement issues, applicants may call the U.S. Government hotline at 1-866-999-HALT (1-866-999-4258).

LICENSE FOR FOREIGN FILING UNDER

Title 35, United States Code, Section 184

Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CFR 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under 37 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Bureau of Industry and Security, Department of Commerce (15 CFR parts 730-774); the Office of Foreign AssetsControl, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

IPR2025-01398
CrowdStrike EX1002 Page 65

SelectUSA

The United States represents the largest, most dynamic marketplace in the world and is an unparalleled location for business investment, innovation, and commercialization of new technologies. The U.S. offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to promote and facilitate business investment. SelectUSA provides information assistance to the international investor community; serves as an ombudsman for existing and potential investors; advocates on behalf of U.S. cities, states, and regions competing for global investment; and counsels U.S. economic development organizations on investment attraction best practices. To learn more about why the United States is the best country in the world to develop technology, manufacture products, deliver services, and grow your business, visit http://www.SelectUSA.gov or call +1-202-482-6800.

IPR2025-01398
CrowdStrike EX1002 Page 66

Doc Code: PA..
Document Description: Power of Attorney

PTO/AIA/82A (07-13)
Approved for use through 03/31/2021. OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

TRANSMITTAL FOR POWER OF ATTORNEY TO ONE OR MORE REGISTERED PRACTITIONERS

NOTE: This form is to be submitted with the Power of Attorney by Applicant form (PTO/AIA/82B) to identify the application to which the Power of Attorney is directed, in accordance with 37 CFR 1.5, unless the application number and filing date are identified in the Power of Attorney by Applicant form. If neither form PTO/AIA/82A nor form PTO/AIA82B identifies the application to which the Power of Attorney is directed, the Power of Attorney will not be recognized in the application.

Application Number	16/405,612
Filing Date	05/07/2019
First Named Inventor	Revanth Patil
Title	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARKWEB
Art Unit	TBA
Examiner Name	TBA
Attorney Docket Number	055743-621457 (M18-191P)

SIGNATURE of Applicant or Patent Practitioner

Signature	/Ari. M. Bai/	Date (Optional)	2019-07-18
Name	Ari M. Bai	Registration Number	
Title (if Applicant is a juristic entity)			
Applicant Name (if Applicant is a juristic entity)			

NOTE: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4(d) for signature requirements and certifications. If more than one applicant, use multiple forms.

☑ *Total of 1 of 5 forms are submitted.

This collection of information is required by 37 CFR 1.131, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Doc Code: PA..
Document Description: Power of Attorney

PTO/AIA/82B (07-13)
Approved for use through 03/31/2021. OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number

POWER OF ATTORNEY BY APPLICANT

I hereby revoke all previous powers of attorney given in the application identified in either the attached transmittal letter or the boxes below.

Application Number	Filing Date
16/405,612	05/07/2019

(Note: The boxes above may be left blank if information is provided on form PTO/AIA/82A.)

☑ I hereby appoint the Patent Practitioner(s) associated with the following Customer Number as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above: 27148

OR

☐ I hereby appoint Practitioner(s) named in the attached list (form PTO/AIA/82C) as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the patent application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above. (Note: Complete form PTO/AIA/82C.)

Please recognize or change the correspondence address for the application identified in the attached transmittal letter or the boxes above to:

☑ The address associated with the above-mentioned Customer Number

OR

☐ The address associated with Customer Number:

OR

☐ Firm or Individual Name					
Address					
City		State		Zip	
Country					
Telephone		Email			

I am the Applicant (If the Applicant is a juristic entity, list the Applicant name in the box):

☑ Inventor or Joint Inventor (title not required below)

☐ Legal Representative of a Deceased or Legally Incapacitated Inventor (title not required below)

☐ Assignee or Person to Whom the Inventor is Under an Obligation to Assign (provide signer's title if applicant is a juristic entity)

☐ Person Who Otherwise Shows Sufficient Proprietary Interest (e.g., a petition under 37 CFR 1.46(b)(2) was granted in the application or is concurrently being filed with this document) (provide signer's title if applicant is a juristic entity)

SIGNATURE of Applicant for Patent

The undersigned (whose title is supplied below) is authorized to act on behalf of the applicant (e.g., where the applicant is a juristic entity).

Signature	Revanth P.	Date (Optional)	
Name	Revanth Patil		
Title	Inventor		

NOTE: Signature - This form must be signed by the applicant in accordance with 37 CFR 1.33. See 37 CFR 1.4 for signature requirements and certifications. If more than one applicant, use multiple forms.

☐ Total of forms are submitted.

This collection of information is required by 37 CFR 1.131, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Doc Code: PA..
Document Description: Power of Attorney

PTO/AIA/82B (07-13)
Approved for use through 03/31/2021. OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number

POWER OF ATTORNEY BY APPLICANT

I hereby revoke all previous powers of attorney given in the application identified in either the attached transmittal letter or the boxes below.

Application Number	Filing Date
16/405,612	05/07/2019

(Note: The boxes above may be left blank if information is provided on form PTO/AIA/82A.)

☑ I hereby appoint the Patent Practitioner(s) associated with the following Customer Number as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above: 27148

OR

☐ I hereby appoint Practitioner(s) named in the attached list (form PTO/AIA/82C) as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the patent application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above. (Note: Complete form PTO/AIA/82C.)

Please recognize or change the correspondence address for the application identified in the attached transmittal letter or the boxes above to:

☑ The address associated with the above-mentioned Customer Number

OR

☐ The address associated with Customer Number:

OR

☐ Firm or Individual Name					
Address					
City		State		Zip	
Country					
Telephone		Email			

I am the Applicant (if the Applicant is a juristic entity, list the Applicant name in the box):

☑ Inventor or Joint Inventor (title not required below)

☐ Legal Representative of a Deceased or Legally Incapacitated Inventor (title not required below)

☐ Assignee or Person to Whom the Inventor is Under an Obligation to Assign (provide signer's title if applicant is a juristic entity)

☐ Person Who Otherwise Shows Sufficient Proprietary Interest (e.g., a petition under 37 CFR 1.46(b)(2) was granted in the application or is concurrently being filed with this document) (provide signer's title if applicant is a juristic entity)

SIGNATURE of Applicant for Patent

The undersigned (whose title is supplied below) is authorized to act on behalf of the applicant (e.g., where the applicant is a juristic entity).

Signature		Date (Optional)	
Name	Paulo Shakarian		
Title	Inventor		

NOTE: Signature - This form must be signed by the applicant in accordance with 37 CFR 1.33. See 37 CFR 1.4 for signature requirements and certifications. If more than one applicant, use multiple forms.

☐ Total of forms are submitted.

This collection of information is required by 37 CFR 1.131, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Doc Code: PA..
Document Description: Power of Attorney

PTO/AIA/82B (07-13)
Approved for use through 03/31/2021. OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number

POWER OF ATTORNEY BY APPLICANT

I hereby revoke all previous powers of attorney given in the application identified in either the attached transmittal letter or the boxes below.

Application Number	Filing Date
16/405,612	05/07/2019

(Note: The boxes above may be left blank if information is provided on form PTO/AIA/82A.)

☑ I hereby appoint the Patent Practitioner(s) associated with the following Customer Number as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above: 27148

OR

☐ I hereby appoint Practitioner(s) named in the attached list (form PTO/AIA/82C) as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the patent application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above. (Note: Complete form PTO/AIA/82C.)

Please recognize or change the correspondence address for the application identified in the attached transmittal letter or the boxes above to:

☑ The address associated with the above-mentioned Customer Number

OR

☐ The address associated with Customer Number:

OR

☐ Firm or Individual Name					
Address					
City		State		Zip	
Country					
Telephone		Email			

I am the Applicant (if the Applicant is a juristic entity, list the Applicant name in the box):

☑ Inventor or Joint Inventor (title not required below)

☐ Legal Representative of a Deceased or Legally Incapacitated Inventor (title not required below)

☐ Assignee or Person to Whom the Inventor is Under an Obligation to Assign (provide signer's title if applicant is a juristic entity)

☐ Person Who Otherwise Shows Sufficient Proprietary Interest (e.g., a petition under 37 CFR 1.46(b)(2) was granted in the application or is concurrently being filed with this document) (provide signer's title if applicant is a juristic entity)

SIGNATURE of Applicant for Patent

The undersigned (whose title is supplied below) is authorized to act on behalf of the applicant (e.g., where the applicant is a juristic entity).

Signature	Ashkan Alaali	Date (Optional)	
Name	Ashkan Alaali		
Title	Inventor		

NOTE: Signature - This form must be signed by the applicant in accordance with 37 CFR 1.33. See 37 CFR 1.4 for signature requirements and certifications. If more than one applicant, use multiple forms.

☐ Total of forms are submitted.

This collection of information is required by 37 CFR 1.131, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Doc Code: PA..
Document Description: Power of Attorney

PTO/AIA/82B (07-13)
Approved for use through 03/31/2021. OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number

POWER OF ATTORNEY BY APPLICANT

I hereby revoke all previous powers of attorney given in the application identified in either the attached transmittal letter or the boxes below.

Application Number	Filing Date
16/405,612	05/07/2019

(Note: The boxes above may be left blank if information is provided on form PTO/AIA/82A.)

☑ I hereby appoint the Patent Practitioner(s) associated with the following Customer Number as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above: 27148

OR

☐ I hereby appoint Practitioner(s) named in the attached list (form PTO/AIA/82C) as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the patent application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above. (Note: Complete form PTO/AIA/82C.)

Please recognize or change the correspondence address for the application identified in the attached transmittal letter or the boxes above to:

☑ The address associated with the above-mentioned Customer Number

OR

☐ The address associated with Customer Number:

OR

☐ Firm or Individual Name					
Address					
City		State		Zip	
Country					
Telephone		Email			

I am the Applicant (if the Applicant is a juristic entity, list the Applicant name in the box):

☑ Inventor or Joint Inventor (title not required below)

☐ Legal Representative of a Deceased or Legally Incapacitated Inventor (title not required below)

☐ Assignee or Person to Whom the Inventor is Under an Obligation to Assign (provide signer's title if applicant is a juristic entity)

☐ Person Who Otherwise Shows Sufficient Proprietary Interest (e.g., a petition under 37 CFR 1.46(b)(2) was granted in the application or is concurrently being filed with this document) (provide signer's title if applicant is a juristic entity)

SIGNATURE of Applicant for Patent

The undersigned (whose title is supplied below) is authorized to act on behalf of the applicant (e.g., where the applicant is a juristic entity).

Signature	Ericsson Santana Marin	Date (Optional)	05/23/2019
Name	Ericsson Marin		
Title	Inventor		

NOTE: Signature - This form must be signed by the applicant in accordance with 37 CFR 1.33. See 37 CFR 1.4 for signature requirements and certifications. If more than one applicant, use multiple forms.

☐ Total of forms are submitted.

This collection of information is required by 37 CFR 1.131, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PTO/AIA/01 (06-12)
Approved for use through 11/30/2020. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB

As the below named inventor, I hereby declare that:

This declaration is directed to:

☐ The attached application, or

☒ United States application or PCT international application number 16/405,612

filed on 05/07/2019

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Revanth Patil Date (Optional): _____

Signature: Revanth P.

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PTO/AIA/01 (06-12)
Approved for use through 11/30/2020. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB

As the below named inventor, I hereby declare that:

This declaration is directed to:

☐ The attached application, or

☒ United States application or PCT international application number 16/405,612

filed on 05/07/2019

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Paulo Shakarian Date (Optional):

Signature:

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PTO/AIA/01 (06-12)
Approved for use through 11/30/2020. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB

As the below named inventor, I hereby declare that:

This declaration is directed to:

☐ The attached application, or

☒ United States application or PCT international application number 16/405,612

filed on 05/07/2019.

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Ashkan Aleali Date (Optional) :

Signature: Ashkan Aleali

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PTO/AIA/01 (06-12)
Approved for use through 11/30/2020. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB

As the below named inventor, I hereby declare that:

This declaration is directed to:

☐ The attached application, or

☒ United States application or PCT international application number 16/405,612

filed on 05/07/2019

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Ericsson Marin Date (Optional): 05/23/2019

Signature: Ericsson Santana Marin

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Electronic Acknowledgement Receipt

EFS ID:	36627598
Application Number:	16405612
International Application Number:	
Confirmation Number:	4562
Title of Invention:	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB
First Named Inventor/Applicant Name:	Revanth Patil
Customer Number:	27148
Filer:	Ari M. Bai/Rebecca Endsley
Filer Authorized By:	Ari M. Bai
Attorney Docket Number:	055743-621457 (M18-191P)
Receipt Date:	18-JUL-2019
Filing Date:	07-MAY-2019
Time Stamp:	16:11:04
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Power of Attorney	621457_Transmittal_of_POA.pdf	360727 d3ff6580e3b86ef40925edca0c556135574f8566	no	1

Warnings:

IPR2025-01398
CrowdStrike EX1002 Page 76

Information:					
2	Power of Attorney	621457_POA.pdf	425061 b2e57a17a52c95bc592737560ddd4690a29fe3ab	no	4
Warnings:					
Information:					
3	Oath or Declaration filed	621457_Declarations.pdf	322079 97522bbae4d252580b420169dbf817988c23f978	no	4
Warnings:					
Information:					
			Total Files Size (in bytes):	1107867	

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

<u>New Applications Under 35 U.S.C. 111</u>
If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.
<u>National Stage of an International Application under 35 U.S.C. 371</u>
If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.
<u>New International Application Filed with the USPTO as a Receiving Office</u>
If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
16/405,612	05/07/2019	Revanth Patil	055743-621457 (M18-191P)

CONFIRMATION NO. 4562

27148
POLSINELLI PC
900 WEST 48TH PLACE
SUITE 900
KANSAS CITY, MO 64112-1895

POA ACCEPTANCE LETTER

OC000000109814277

Date Mailed: 07/23/2019

NOTICE OF ACCEPTANCE OF POWER OF ATTORNEY

This is in response to the Power of Attorney filed 07/18/2019.

The Power of Attorney in this application is accepted. Correspondence in this application will be mailed to the above address as provided by 37 CFR 1.33.

Questions about the contents of this notice and the requirements it sets forth should be directed to the Office of Data Management, Application Assistance Unit, at **(571) 272-4000** or **(571) 272-4200** or **1-888-786-0101**.

/zabraha/

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
16/405,612	05/07/2019	Revanth Patil	055743-621457 (M18-191P)

CONFIRMATION NO. 4562

27148
POLSINELLI PC
900 WEST 48TH PLACE
SUITE 900
KANSAS CITY, MO 64112-1895

PUBLICATION NOTICE

OC000000112625479

Title:SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB

Publication No.US-2019-0347327-A1
Publication Date:11/14/2019

NOTICE OF PUBLICATION OF APPLICATION

The above-identified application will be electronically published as a patent application publication pursuant to 37 CFR 1.211, et seq. The patent application publication number and publication date are set forth above.

The publication may be accessed through the USPTO's publically available Searchable Databases via the Internet at www.uspto.gov. The direct link to access the publication is currently http://www.uspto.gov/patft/.

The publication process established by the Office does not provide for mailing a copy of the publication to applicant. A copy of the publication may be obtained from the Office upon payment of the appropriate fee set forth in 37 CFR 1.19(a)(1). Orders for copies of patent application publications are handled by the USPTO's Public Records Division. The Public Records Division can be reached by telephone at (571) 272-3150 or (800) 972-6382, by facsimile at (571) 273-3250, by mail addressed to the United States Patent and Trademark Office, Public Records Division, Alexandria, VA 22313-1450 or via the Internet.

In addition, information on the status of the application, including the mailing date of Office actions and the dates of receipt of correspondence filed in the Office, may also be accessed via the Internet through the Patent Electronic Business Center at www.uspto.gov using the public side of the Patent Application Information and Retrieval (PAIR) system. The direct link to access this status information is currently https://portal.uspto.gov/pair/PublicPair. Prior to publication, such status information is confidential and may only be obtained by applicant using the private side of PAIR.

Further assistance in electronically accessing the publication, or about PAIR, is available by calling the Patent Electronic Business Center at 1-866-217-9197.

Office of Data Managment, Application Assistance Unit (571) 272-4000, or (571) 272-4200, or 1-888-786-0101

IPR2025-01398
CrowdStrike EX1002 Page 79

PTO/SB/08a (03-15)
Approved for use through 07/31/2016. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT

(Not for submission under 37 CFR 1.99)

Application Number	16405612
Filing Date	2019-05-07
First Named Inventor	Revanth Patil
Art Unit	2658
Examiner Name	Richemond Dorvil
Attorney Docket Number	055743-621457 M18-191P

U.S.PATENTS

Examiner Initial*	Cite No	Patent Number	Kind Code[1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1	10176438	B2	2019-01-08	Shakarian et al.	
	2	10313385	B2	2019-06-04	Shakarian et al.	
	3	10437945	B2	2019-10-08	Shakarian et al.	

If you wish to add additional U.S. Patent citation information please click the Add button.

U.S.PATENT APPLICATION PUBLICATIONS

Examiner Initial*	Cite No	Publication Number	Kind Code[1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1	20190349393	A1	2019-11-14	Nunes et al.	
	2	20200036743	A1	2020-01-30	Almukaynizi et al.	

If you wish to add additional U.S. Published Application citation information please click the Add button.

FOREIGN PATENT DOCUMENTS

Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2]	Kind Code[4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear	T[5]

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16405612
	Filing Date	2019-05-07
	First Named Inventor	Revanth Patil
	Art Unit	2658
	Examiner Name	Richemond Dorvil
	Attorney Docket Number	055743-621457 M18-191P

	1	2019089389	WO	A1	2019-05-09	Shakarian et al.		☐
	2	2019157335	WO	A1	2019-08-15	Shaabani et al.		☐

If you wish to add additional Foreign Patent Document citation information please click the Add button

NON-PATENT LITERATURE DOCUMENTS

Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc), date, pages(s), volume-issue number(s), publisher, city and/or country where published.	T5
	1	U.S. Application No. 16/424,929, Shakarian et al., filed 2016-11-28.	☐
	2	U.S. Application No. 16/548,329, Tavabi et al., filed 2019-08-22.	☐
	3	U.S. Application No. 16/640,878, Shakarian et al., filed 2018-10-26.	☐
	4	U.S. Application No. 16/653,899, Sarkar et al., filed 2019-10-15.	☐
	5	U.S. Application No. 16/875,540, Nunes et al., filed 2020-05-15.	☐
	6	Tor Project, "Tor: Overview" [Online] Available: https://www.torproject.org/about/overview.html.en , accessed 27-December-2018" ("2017")	☐
	7	Abbasi, et al., "Descriptive analytics: Examining expert hackers in web forums", Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint, pp. 56–63 (IEEE, 2014).	☐

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16405612
	Filing Date	2019-05-07
	First Named Inventor	Revanth Patil
	Art Unit	2658
	Examiner Name	Richemond Dorvil
	Attorney Docket Number	055743-621457 M18-191P

	8	Breiman, et al., "Random forests-classification description", Department of Statistics, Berkeley 2 (2007).	☐
	9	Chawla, et al., "Smote: synthetic minority over-sampling technique", Journal of artificial intelligence research 16, 321–357 (2002).	☐
	10	Chertoff, et al., "The impact of the dark web on internet governance and cyber security", (2015).	☐
	11	Dekel, et al., "Multiclass-multilabel classification with more classes than examples", Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 137–144 (2010).	☐
	12	Harris, "Distributional structure", Word 10, 2-3, 146–162 (1954).	☐
	13	Le, et al., "Distributed representations of sentences and documents", Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 1188–1196 (2014).	☐
	14	Manning, "Introduction", in Retrieving the Radical Tillich, pp. 1–19 (Springer, 2015).	☐
	15	Marin, et al., "Product offerings in malicious hacker markets", Intelligence and Security Informatics (ISI), 2016 IEEE Conference, pp. 187– 189 (IEEE, 2016).	☐
	16	Moore, et al., "Cryptopolitik and the darknet", Survival 58, 1, 7–38 (2016).	☐
	17	Nunes, et al., "Darknet and deepnet mining for proactive cybersecurity threat intelligence", Intelligence and Security Informatics (ISI), 2016 IEEE Conference, pp. 7–12 (IEEE, 2016).	☐
	18	Okapi-BM25, "Okapi-bm25 ranking function Wikipedia, the free encyclopedia", URL https://en.wikipedia.org/wiki/Okapi-BM25, [Online; accessed 12-December-2017] (2017).	☐

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	16405612
	Filing Date	2019-05-07
	First Named Inventor	Revanth Patil
	Art Unit	2658
	Examiner Name	Richemond Dorvil
	Attorney Docket Number	055743-621457 M18-191P

	19	Pedregosa, et al., "Scikit-learn: Machine learning in Python", Journal of Machine Learning Research 12, 2825–2830 (2011).	☐
	20	Pérez-Iglesias, et al., "Integrating the probabilistic models bm25/bm25f into lucene", arXiv preprint arXiv:0911.5046 (2009).	☐
	21	Rajaraman, Jd (2011). Data mining, Mining of Massive Datasets pp. 1–17 (2011).	☐
	22	Rong, "word2vec parameter learning explained", arXiv preprint arXiv:1411.2738 (2014).	☐
	23	Samtani, et al., "Azsecure hacker assets portal: Cyber threat intelligence and malware analysis", Intelligence and Security Informatics (ISI), 2016 IEEE Conference, pp. 19–24 (IEEE, 2016).	☐
	24	Sanderson, Introduction to information retrieval, cambridge university press 2009, Natural Language Engineering 16, 1, 232–234 (2010).	☐
	25	Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval", Journal of documentation 28, 1, 11–21 (1972).	☐
	26	Wikipedia, "Random forest — wikipedia, the free encyclopedia", URL https://en.wikipedia.org/w/index.php?title=Randomforestoldid= 822509370, [Online; accessed29 − December − 2018].	☐

If you wish to add additional non-patent literature document citation information please click the Add button

EXAMINER SIGNATURE

Examiner Signature		Date Considered	

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document. [4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT

(Not for submission under 37 CFR 1.99)

Application Number	16405612
Filing Date	2019-05-07
First Named Inventor	Revanth Patil
Art Unit	2658
Examiner Name	Richemond Dorvil
Attorney Docket Number	055743-621457 M18-191P

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

☐ That each item of information contained in the information disclosure statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☐ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

☐ See attached certification statement.

☐ The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

☒ A certification statement is not submitted herewith.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	/Ari M. Bai/	Date (YYYY-MM-DD)	2020-08-17
Name/Print	Ari M. Bai	Registration Number	38726

This collection of information is required by 37 CFR 1.97 and 1.98. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 1 hour to complete, including gathering, preparing and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these record s.

2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.

3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.

4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).

5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.

6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).

7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.

8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.

9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
09 May 2019 (09.05.2019)

WIPO | PCT

(10) International Publication Number
WO 2019/089389 A1

(51) International Patent Classification:
G06F 11/00 (2006.01) *G06F 12/14* (2006.01)
G06F 12/16 (2006.01)

(21) International Application Number:
PCT/US2018/057812

(22) International Filing Date:
26 October 2018 (26.10.2018)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/581,123 03 November 2017 (03.11.2017) US

(71) Applicant *(for all designated States except US)*: **ARIZONA BOARD OF REGENTS ON BEHALF OF ARIZONA STATE UNIVERSITY** [US/US]; 699 S. Mills Avenue, Suite 601, Room 691AA, Tempe, Arizona 85281 (US).

(72) Inventors; and
(71) Applicants *(for US only)*: **SHAKARIAN, Paulo** [US/US]; c/o ARIZONA BOARD OF REGENTS ON BEHALF OF ARIZONA STATE UNIVERSITY, 699 S. Mills Avenue, Suite 601, Room 691AA, Tempe, Arizona 85281 (US). **ALMUKAYNIZI, Mohammed** [SA/US]; c/o ARIZONA BOARD OF REGENTS ON BEHALF OF ARIZONA STATE UNIVERSITY, 699 S. Mills Avenue, Suite 601, Room 691AA, Tempe, Arizona 85281 (US). **SHAKARIAN, Jana** [US/US]; c/o ARIZONA BOARD OF REGENTS ON BEHALF OF ARIZONA STATE UNIVERSITY, 699 S. Mills Avenue, Suite 601, Room 691AA, Tempe, Arizona 85281 (US). **NUNES, Eric** [IN/US]; c/o ARIZONA BOARD OF REGENTS ON BEHALF OF ARIZONA STATE UNIVERSITY, 699 S. Mills Avenue, Suite 601, Room 691AA, Tempe, Arizona 85281 (US). **DHARAIYA, Krishna** [IN/US]; c/o ARIZONA BOARD OF REGENTS ON BEHALF OF ARIZONA STATE UNIVERSITY, 699 S. Mills Avenue, Suite 601, Room 691AA, Tempe, Arizona 85281 (US). **SENGUTTUVAN, Manoj Balasubramaniam** [IN/US]; c/o ARIZONA BOARD OF REGENTS ON BEHALF OF ARIZONA STATE UNIVERSITY, 699 S. Mills Avenue, Suite 601, Room 691AA, Tempe, Arizona 85281 (US). **GRIMM, Alexander** [US/US]; c/o ARIZONA BOARD OF REGENTS ON BEHALF OF ARIZONA STATE UNIVERSITY, 699 S. Mills Avenue, Suite 601, Room 691AA, Tempe, Arizona 85281 (US).

(54) **Title:** SYSTEMS AND METHODS FOR PRIORITIZING SOFTWARE VULNERABILITIES FOR PATCHING

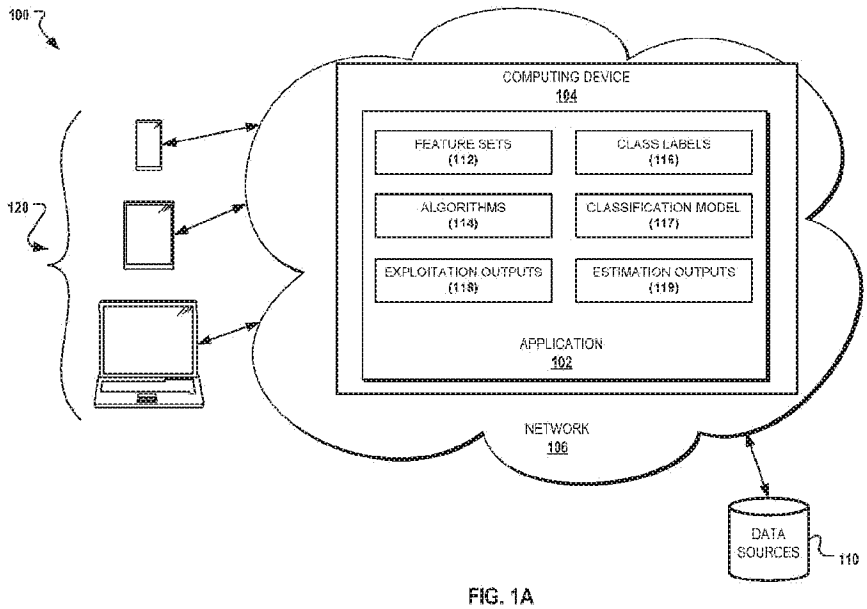


FIG. 1A

(57) **Abstract:** Various embodiments for predicting which software vulnerabilities will be exploited by malicious hackers and hence prioritized by patching are disclosed.

WO 2019/089389 A1

[Continued on next page]

IPR2025-01398
CrowdStrike EX1002 Page 86

(74) **Agent: BAI, Ari** et al.; POLSINELLI PC, 100 South Fourth Street, Suite 1000, St. Louis, Missouri 63102-1825 (US).

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

SYSTEMS AND METHODS FOR PRIORITIZING SOFTWARE VULNERABILITIES FOR PATCHING

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This is a PCT application that claims benefit to U.S. provisional application serial number 62/581,123 filed on November 3, 2017 which is incorporated by reference in its entirety.

GOVERNMENT SUPPORT

[0002] This invention was made with government support under Contract No. N00014-15-1-2742 from the Office of Naval Research and Contract No. FA8750-16-C-0012 from the Intelligence Advanced Research Projects Activity (IARPA). The Government has certain rights in the invention.

FIELD

[0003] The present disclosure generally relates to assessing the likelihood of exploitation of software vulnerabilities, and in particular to systems and methods for predicting which software vulnerabilities will be exploited by malicious hackers and hence prioritized by patching.

BACKGROUND

[0004] An increasing number of software vulnerabilities are discovered and publicly disclosed every year. In 2016 alone, more than 10,000 vulnerability identifiers were assigned and at least 6,000 were publicly disclosed by the National Institute of Standards and Technology (NIST). Once the vulnerabilities are disclosed publicly, the likelihood of those vulnerabilities being exploited increases. With limited resources, organizations often look to prioritize which vulnerabilities to patch by assessing the impact it will have on the organization if exploited. Standard risk *assessment systems* such as Common Vulnerability Scoring System (CVSS), Microsoft Exploitability Index, Adobe Priority Rating report many vulnerabilities as severe and will be exploited to be on the side of caution. This does not alleviate the problem much since the majority of the flagged vulnerabilities will not be attacked.

[0005] Further, current methods for prioritizing patching vulnerabilities appear to fall short. Verizon reported that over 99% of breaches are caused by

IPR2025-01398
CrowdStrike EX1002 Page 88

exploits to known vulnerabilities. Cisco also reported that "The gap between the availability and the actual implementation of such patches is giving attackers an opportunity to launch exploits." For some vulnerabilities, the time window to patch the system is very small. For instance, exploits targeting the Heartbleed bug in OpenSSL were detected in the wild 21 hours after the vulnerability was publicly disclosed. Hence, organizations need to efficiently assess vulnerabilities that will be exploited in the wild while keeping the false alarm rate low.

[0006] NIST provides the National Vulnerability Database (NVD) which comprises of a comprehensive list of vulnerabilities disclosed, but only a small fraction of those vulnerabilities (less than 3%) are found to be exploited in the wild - a result confirmed in the present disclosure. Further, previous work has found that the CVSS score provided by NIST is not an effective predictor of vulnerabilities being exploited. It has previously been proposed that other methods such as the use of social media, darkweb markets and certain white-hat websites would be suitable alternatives. However, this previous work has its limitations. For instance, methodical concerns on the use of social media for exploit prediction were recently raised; data feeds for proactive exploit prediction were limited to single sites that provided a relatively small number of predictions; and other work demonstrates the viability of data collection but does not quantify the results of prediction.

[0007] It is with these observations in mind, among others, that various aspects of the present disclosure were conceived and developed.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] FIG. **1A** is a network diagram illustrating an exemplary computer-implemented system for implementing vulnerability prediction as described herein, according to aspects of the present disclosure;

[0009] FIG. **1B** is a simplified block diagram showing a first embodiment (embodiment A) of the exploit prediction model, according to aspects of the present disclosure;

[0010] FIG. **2** is a graphical representation showing the vulnerabilities disclosed per month, according to data examined by embodiment A and aspects of the present disclosure;

[0011] FIG. **3** is a graphical representation of the day difference between Common Vulnerabilities and Exposures (CVE) first published in the

IPR2025-01398
CrowdStrike EX1002 Page 89

National Vulnerability Database (NVD) and Symantec attack signature date vs. the fraction of exploited CVE's on NVD reported, according to data examined by embodiment A and aspects of the present disclosure;

[0012] FIG. **4** is a graphical representation of a day difference between proof-of-concept CVE first mentioned and Symantec attack signature date vs. the fraction of exploited CVE's on ExploitDB (EDB) reported, according to data examined by embodiment A and aspects of the present disclosure;

[0013] FIG. **5** is a graphical representation of the day difference between CVE first mentioned in darkweb and deepweb and Symantec attack signature date vs. the fraction of exploited CVE's on darkweb and deepweb reported, according to data examined by embodiment A and aspects of the present disclosure;

[0014] FIG. **6** is a graphical representation of the most exploited vendors, according to data examined by embodiment A and aspects of the present disclosure;

[0015] FIG. **7** is a graphical representation showing the percentage of vulnerabilities mentions in each language and the percentage of the exploited vulnerabilities mentioned in each language, according to data examined by embodiment A and aspects of the present disclosure;

[0016] FIG. **8** is a graphical representation of a precision-recall curve for proposed features for Microsoft-Adobe vulnerabilities, according to data examined by embodiment A and aspects of the present disclosure;

[0017] FIG. **9** is a graphical representation of a precision and recall for classification based on CVSS base score version 2.0 threshold, according to data examined by embodiment A and aspects of the present disclosure;

[0018] FIG. **10** is a graphical representation of a precision-recall curve for classification based on CVSS score threshold (RF), according to data examined by embodiment A and aspects of the present disclosure;

[0019] FIG. **11** is a graphical representation of a ROC curve for classification based on Random Forest classifier, according to data examined by embodiment A and aspects of the present disclosure;

[0020] FIGS. **12A** and **12B** are graphical representations of an ROC curve using strategy 1 and strategy 2, respectively, according to data examined by embodiment A and aspects of the present disclosure;

IPR2025-01398
CrowdStrike EX1002 Page 90

[0021] FIG. **13** is an overview of the prediction model, according to aspects of the present disclosure;

[0022] FIG. **14** is a simplified block diagram that illustrates a second embodiment (embodiment B) of the exploit prediction model, according to aspects of the present disclosure;

[0023] FIG. **15** is a histogram showing vulnerabilities represented per year from NVD, DW and Symantec, according to aspects of the present disclosure;

[0024] FIG. **16** is a graphical representation showing an ablation test, according to aspects of the present disclosure;

[0025] FIG. **17** is a graphical representation showing test results on individual feature sets, according to aspects of the present disclosure; and

[0026] FIG. **18** is a simplified block diagram showing an example of a computing system that may implement various services, systems, and methods discussed herein.

[0027] Corresponding reference characters indicate corresponding elements among the view of the drawings. The headings used in the figures do not limit the scope of the claims.

DETAILED DESCRIPTION

[0028] In the present disclosure, the ability to predict exploits in the wild is examined using cyber threat intelligence feeds comprised from a variety of data sources or data feeds. This problem is of direct relevance to patch prioritization. Referencing data gathered or otherwise accessed from darkweb and deepweb (DW) and conducting much analysis on data feeds collected from various online sources (e.g., SecurityFocus, Talos), and after over one hundred interviews with professionals working for managed security service providers (MSSP's), firms specializing in cyber risk assessment, and security specialists working for managed (IT) service providers (MSP's), many sources of data were identified that were representative of current threat intelligence used for vulnerability prioritization – examples being presented in FIG. **13**. Amongst those, three are examined in embodiment A: (1) ExploitDB (EDB) contains information on proof-of-concepts for vulnerabilities provided by security researchers from various blogs and security reports, (2) Zero Day Initiative (ZDI) is curated by a commercial firm called TippingPoint and uses a variety of reported sources focused on disclosures by

IPR2025-01398
CrowdStrike EX1002 Page 91

various software vendors and their security researchers, and (3) a collection of information scraped from over 120 sites on the darkweb and deepweb (DW) sites. The intuition behind each of these feeds was to not only utilize information that was aggregated over numerous related sources, but also to represent feeds commonly used by cybersecurity professionals.

[0029] The present disclosure demonstrates the utility of the developed machine learning models described herein in predicting exploits in the wild with True Positive Rate (TPR) of 90% while maintaining the False Positive Rate (FPR) less than 15% for a sample data examined by embodiment A. The present disclosure also compares the present system embodiment A to recent benchmarks for exploit prediction, demonstrating the achievement of significant high precision while maintaining the recall under the assumptions made in the present disclosure. The performance of variants of the present embodiment A was also examined in the case when temporal mixing is controlled, and in the case where only a single source is used. Robustness against adversarial data manipulation is also discussed.

[0030] Using vulnerability mentions on EDB, ZDI, and DW, the present disclosure further describes the increase in the vulnerability exploitation likelihood over vulnerabilities only disclosed on NVD. In addition, the present disclosure provides results that demonstrate the likelihood of being exploited given vulnerability mention on EDB is 9%, ZDI is 12% and DW is 14% as compared to NVD, which is 2.4%, and explains the availability of such information relative to the time an exploit is found in the wild.

[0031] The present disclosure also analyzes exploited vulnerabilities based on various other features derived from these data sources such as language used. It is specifically noted that Russian language sites on the darkweb discuss vulnerabilities that are 19 times more likely to be exploited than random, which was greater than other languages examined. Further research was made regarding the probability of exploitation with regard to both data sources and the vendor of the vulnerable software.

[0032] The present disclosure also presents an embodiment (embodiment B) that leverages machine learning techniques on features derived from the social network of users participating in darkweb/deepweb (DW) forums, as well as features derived from the National Vulnerability Database. It specifically demonstrates the viability of user connectivity metadata in predicting the likelihood of

IPR2025-01398
CrowdStrike EX1002 Page 92

vulnerability exploitation with an F1 measure of 0.67 for a subset of software vulnerabilities mentioned in DW.

Vulnerability and Exploits

[0033] Vulnerability is a weakness in a software system that can be exploited by an attacker to compromise the confidentiality, integrity or availability of the system to cause harm. The National Institute of Standards and Technology (NIST) maintains a comprehensive list of publicly disclosed vulnerabilities in its National Vulnerability Database (NVD). NVD also provides information regarding the target softwares (CPE), severity rating (CVSS) in terms of exploitability and impact and the date the vulnerability was published.

[0034] An *exploit* is defined as a piece of code that modifies the functionality of a system using an existing vulnerability. We term the exploits that have been used to target systems in real-world by an attacker as *real-world exploits*. On the other hand, *proof-of-concept exploits* are developed to verify a disclosed vulnerability and might require addition of functionalities to be used in a real-world attack. Proof-of-concept exploits that are identified in the wild (real-world attacks) fall under real-world exploits. Although the chances of detecting real-world exploits if proof-of-concept is already present is high since attackers have a template assisting on implementing exploits, presence of proof-of-concept does not imply that it has been used in the wild.

Technical Challenges

[0035] There are known technical challenges including methodological issues with previous exploit prediction methods and studies. It is also noted that there is a balance between ensuring an evaluation is conducted under real-world conditions and conducting an evaluation on an adequate sample size. A review of some of these challenges is discussed below.

[0036] **Class Imbalance**. As described herein, around 2.4% of the reported vulnerabilities are exploited in real world attacks. This skews the distribution towards one class i.e. *not exploited* in the prediction problem. In such cases, standard machine learning approaches favor the majority class leading to poor performance on the minority class. The present disclosure examines over-sampling techniques (namely SMOTE) only on the training dataset to account for

IPR2025-01398
CrowdStrike EX1002 Page 93

data with sever class imbalance and avoid making the classifier always favor the majority class. Only a marginal improvement is observed for some classifiers, while other classifiers have been negatively affected when tested on unsampled data.

[0037] Evaluating Models on Temporal Data. Machine learning models are evaluated by training the model on one set of data and then testing the model on another set that is assumed to be drawn from the same distribution. The data split can be done randomly or in a stratified manner where the class ratio is maintained in both training and testing. Yet, exploit prediction is a time dependent prediction problem. Hence splitting the data randomly, violates the temporal aspect of the data as events that happen in the future will now be used to predict events that happen in the past, which violates the sequential nature of vulnerability disclosure events. This leads to temporal intermixing of the data. In the results described and related to the embodiment systems presented in this disclosure, this temporal mixing was reduced or avoided in most experiments. However, in certain experiments, where sample size is very small, experiments were included where this is not controlled (as one of our ground-truth sources does not have date/time information). It was explicitly noted when this was the case.

[0038] Vulnerabilities Exploited Before Disclosure. For exploit prediction, the goal is to predict whether a disclosed vulnerability will be exploited in the future or not. Few vulnerabilities are exploited before they are disclosed, hence prediction for such vulnerabilities does not add any value to the goal of the problem. That being said, predicting exploits that have already been used in attacks is important because not all vulnerabilities are patched in a timely manner. Knowing what vulnerabilities are exploited in the wild can help organizations with their cyber defense strategies.

[0039] Limitations of Ground Truth. Attack signatures identified for exploits that were detected in the wild are reported by a well-known cyber defense firm, called Symantec. These are used as ground truth for the exploited vulnerabilities when evaluating the present embodiments described herein. This ground truth is not comprehensive as the distribution of the exploited vulnerabilities over software vendors is found to differ from that for overall vulnerabilities (i.e., vulnerabilities affect products run on Microsoft have well coverage as compared to other OS vendors). Although this source has coverage limitations, it was found to be the most reliable source for exploited vulnerabilities since it reports attack signatures

IPR2025-01398
CrowdStrike EX1002 Page 94

for known vulnerabilities. To avoid over-fitting the machine learning model on this not-so-representative ground truth, the software vendor was omitted from the set of features that were examined in all experiments.

EXPLOIT PREDICTION MODEL

[0040] Predicting the likelihood of vulnerability exploitation through the usage of machine learning techniques has interesting security implications in terms of prioritizing which vulnerabilities need to be patched first to minimize risk of cyberattack. FIG. **13** provides an overview of an exploit prediction model proposed by the present disclosure; while FIG. **1B** gives an overview of a first embodiment A of the exploit prediction model, and FIG. **14** gives an overview of a second embodiment B of the exploit prediction model.

[0041] Referring to FIG. **1A**, an exemplary computer-implemented system (hereinafter "system") **100** for implementing functionality associated with exploit prediction is shown. The system **100** may be utilized to implement the embodiment A described herein, and/or the embodiment B described herein, or other such embodiments or models for exploit prediction. In general, the system **100** describes the configuration of at least one computing device for implementing the functionality associated with the exploit prediction embodiments described herein.

[0042] Specifically, in some embodiments, the system **100** may include and/or generally support functionality defined by an application **102**, defining functionality associated with features of the exploit prediction model described herein. The application **102** may be hosted on or otherwise implemented using one or more of a computing device **104**, which may include a server, controller, a personal computer, a terminal, a workstation, a portable computer, a mobile device, a tablet, a mainframe, or other such computing device.

[0043] By installing, executing, or otherwise implementing the application **102**, the computing device **104** is configured for administering, processing, and providing access to aspects of the application **102** via a network **106**, which may include the Internet, an intranet, a virtual private network (VPN), and the like. In some embodiments, a cloud (not shown) may be implemented to execute one or more components of the system **100**. In addition, aspects of the system **100** and/or the application **102** may be provided using platform as a service (PaaS),

IPR2025-01398
CrowdStrike EX1002 Page 95

and/or software as a service (SaaS) using e.g., Amazon Web Services, or other distributed systems.

[0044] As further indicated in FIG. **1A** and described herein, the computing device **104** and the application **102** may access information from one or more data sources **110** or data feeds. The application **102** may further implement functionality associated with the exploit prediction model defined by various modules; namely, a feature sets **112** module, an algorithms module **114**, a class labels module **116**, a classification model module **117**, an exploitation outputs module **118** and an estimation outputs module **119**. The aforementioned modules are merely exemplary and it is contemplated that the application **102** may define any number of related modules or features for implementing the functionality of the exploit prediction model described herein. Further, at least some features of the application **102** may be made available to a plurality of user devices **120** in communication with the computing device **104** of the system **100** and/or the local computing device **108** via the network **106**.

[0045] The embodiments present in this disclosure consist of the following four phases, illustrated in FIG. **13**:

- **Obtaining Data (202):** Three of the data sources **110** are utilized in addition to NVD in embodiment A of this disclosure. These data sources include EDB (ExploitDB), ZDI (Zero Day Initiative) and data mined from DW (darkweb and deepnet) markets and forums focusing on malicious hacking. On the other hand, only a sample of DW sites may be used to show the viability of features derived from the properties of social network of users discussing vulnerabilities in DW as presented in embodiment B of this disclosure.
- **Computing Features and Class Labels (Generating Classification Models (204)):** Features are extracted from each of the data sources **110**. The features include, but are not limited to, bag of words features for vulnerability description and discussions on the DW, binary features which checks for the presence of proof-of-concept exploits in EDB, vulnerability disclosures in ZDI and DW. Additional features are included from NVD namely, CVSS score, CVSS vector, and CWE. The class labels are determined based on a ground truth data, which comprise a set of attack signatures of exploits detected in

IPR2025-01398
CrowdStrike EX1002 Page 96

the wild, along with description of vulnerabilities leveraged by these exploits. This data may be obtained from Symantec.

- **Training the Classification Models (205)**
- **Computing the Likelihood of Exploitation (206):** In all embodiments presented in this disclosure, the problem is viewed as a binary classification problem, and several standard supervised machine learning approaches were evaluated and may be implemented.
- **Generating Estimation Outputs (208):** Different algorithms can be used to present the results of the likelihood of exploitation. Among which are rescaling the results, ranking vulnerabilities, categorizing vulnerabilities under pre-determine categories.

Data Sources (110)

[0046] The present disclosure contemplates the combination of vulnerability and exploit information from multiple open source databases, namely: The National Vulnerability Database (NVD), The Exploit Database (EDB), Zero Day Initiative (ZDI), Darkweb database collected from mining darknet marketplaces and forums. For experiments related to the described model examined with respect to embodiment A, efforts were concentrated towards vulnerabilities disclosed in 2015 and 2016. Table **1** shows the vulnerabilities identified from each of the data sources between 2015 and 2016 as well as the number of vulnerabilities that were exploited in real-world attacks. The exploited vulnerabilities are used as ground truth for the described prediction experiments. A brief overview of each of the data sources used in embodiment A, including ground truth, is provided.

Table 1: Number of Vulnerabilities (2015-2016)

Database	Vulnerabilities	Exploited	% Exploited
NVD	12598	306	2.4%
EDB	799	74	9.2%
ZDI	824	95	11.5%
DW	378	52	13.8%

[0047] **NVD**. The National Vulnerability Database maintains a database of publicly disclosed vulnerabilities each one identified using a unique

IPR2025-01398
CrowdStrike EX1002 Page 97

CVE-ID. Vulnerabilities were collected as disclosed between 2015 and 2016. The resulting dataset is comprised of 12598 vulnerabilities. FIG. **2** shows the month wise disclosure of vulnerabilities. At the time of data collection there were only 30 vulnerabilities disclosed in December 2016, hence the small bar at the end of 2016. For each vulnerability, its description, CVSS score and vector are gathered. Organizations often use the CVSS score to prioritize which vulnerabilities to patch. The CVSS vector lists the components from which the score is computed. The components of the vector include Access Complexity, Authentication, Confidentiality, Integrity and Availability. Access complexity indicates how difficult is to exploit the vulnerability once the attacker has gained access to the target system. It is defined in terms of three levels: High, Medium and Low. Authentication indicates whether authentication is required by the attacker to exploit the vulnerability. It is a binary identifier taking the values Required and Not Required. Confidentiality, Integrity and Availability indicate what loss the system would incur if the vulnerability is exploited. It takes the values None, Partial and Complete.

[0048] **EDB (white-hat community)**. Exploit Database is a collection of proof-of-concept exploits maintained by Offensive Security Training that has CVE's associated with available exploits. Using the unique CVE-ID's from the NVD database for the time period between 2015 and 2016, the EDB was queried to find whether a proof-of-concept exploit was available. The date the proof-of-concept exploit was also recorded and posted for these experiments. Using the CVE ID's from NVD, we query EDB to find 799 vulnerabilities with verified proof-of-concepts.

[0049] **ZDI (vulnerability detection community)**. Zero Day Initiative launched by TippingPoint maintains a database of vulnerabilities submitted by security researchers. The submitted vulnerability is first verified before it is added to the database. Monetary incentive is provided if the vulnerability is verified to the researcher. ZDI then notifies the vendor to develop patches for the vulnerability before public disclosure. The ZDI database was queried to collect information regarding vulnerabilities that might have been disclosed by ZDI. Between 2015 and 2016, the query returned 824 vulnerabilities common between NVD and ZDI. The disclosed vulnerabilities were also made of note.

[0050] **DW (black-hat community)**. The data collection infrastructure was summarized. In the present disclosure, the exploit prediction model may be implemented to crawl websites on DW, both marketplaces and forums, to collect

IPR2025-01398
CrowdStrike EX1002 Page 98

data relating to malicious hacking. Sites are first identified before developing scripts for automatic data collection. A site is being put forward to script development after it has been established that the content is of interest (hacking-related) and relatively stable. The population size of the site is being observed, though not much decisive power is assigned to it. While a large population is an indicator for the age and stability of the site, a small population number can be associated with higher-value information (closed forums). While it would be incorrect to label forum users as criminals, there are clearly users communicating malicious intent and sometimes malicious code is exchanged.

[0051] Users in DW advertise and sell their wares on marketplaces. Hence, DW marketplaces provide a new avenue to gather information about vulnerabilities and exploits. Forums on the other hand, feature discussions on newly discovered vulnerabilities and exploits kits. Data related to malicious hacking is filtered from the noise and added to a database using a machine learning approach with high precision and recall. Not all exploits or vulnerability items in the database have a CVE number associated with them. First, the subject database may be queried to extract all items with CVE mentions. Some vulnerabilities are mentioned in DW using Microsoft Security Bulletin Number (e.g., MS16-006) every bulletin number was mapped to its corresponding CVE ID, making ground truth assignment easy. These items can be both products sold on markets as well as posts extracted from forums discussing topics relating to malicious hacking. 378 unique CVE mentions were found between 2015 and 2016 from more than 120 DW websites. This number is a lot more than previous works have discovered (n = 103). The posting date and descriptions associated with all the CVE mentions were also queried including product title and description, vendor information, entire discussion with the CVE mention, author of the posts, topic of the discussion.

[0052] Analyzing a sample of the DW items with CVE mentions, it was found that all items are posted with item names (a mandatory field for the majority of marketplaces) while only half of them are with description. The item names are mostly short (median number of words = 7). Additionally, almost all of the items seem to advertise exploits — though in many cases it is implicit, and about 88.5% of these vulnerabilities are not reported by Symantec; although 16% of them target products run on Microsoft. The inventors chose not to label these samples as positives since there is no evidence supporting that these are functional exploits nor

IPR2025-01398
CrowdStrike EX1002 Page 99

were used in the wild. From all items vulnerabilities, only 2% appeared in DW before reported by Symantec while 9.5% are either reported by Symantec without date (7% of the total items vulnerabilities) or with date before they appeared in DW (2.5%). This implies that early predicting what vulnerabilities will be exploited is a non-trivial task when only using events in the past when predicting future. Further discussions on more observations are presented below.

[0053] **Attack Signatures (Ground truth)**. For our ground truth, vulnerabilities that were exploited in the wild were identified using Symantec's anti-virus attack signatures and Intrusion Detection Systems' (IDS) attacks signatures. The attack signatures were associated with the CVE-ID of the vulnerability that was exploited. These CVE's were mapped to the CVE's mined from NVD, EDB, ZDI and DW. This ground truth indicates actual exploits that were used in the wild and not just proof-of-concept exploits used as ground truth in related work. Table **1** shows the number of vulnerabilities exploited as compared to the ones disclosed in 2015 and 2016 for all the data sources considered. For NVD, around 2.4% of the disclosed vulnerabilities were exploited, which is consistent with previous literature. On the other hand for EDB, ZDI and DW there is a significant percent increase in exploited vulnerabilities to 9%, 12% and 14% respectively. This shows that it was more likely to identify a vulnerability that will be exploited in the future if it has a proof-of-concept available in EDB or mentions in ZDI or DW. For this research, there was no data regarding the volume and frequency of the attacks carried by exploits; hence all exploited vulnerabilities were considered to have equal importance and deserve the same amount of consideration. Additionally, the *exploitation date* of a vulnerability was defined as the date it was first detected in the wild. Symantec IDS attack signatures are reported without recoding the dates when they were first detected, but anti-virus attacks signatures are reported with their *exploitation date*. Within 2015 and 2016, the attack signatures reported without dates are 112 while those reported with their discovery dates are 194.

[0054] The second embodiment of the present disclosure (embodiment B) is evaluated on a database of 46 darkweb and deepweb forums collected from the same data collection infrastructure used by embodiment A. However, a longer time period of data is considered in embodiment B. All postings between January 2010 and March 2017 are analyzed or otherwise considered. FIG. **15** is a histogram

IPR2025-01398
CrowdStrike EX1002 Page 100

showing the vulnerability reported per year from NVD, DW, and Symantec. Additionally, the same ground truth data is used to label exploited vulnerabilities.

Feature Description

[0055] Features used by embodiment A were combined from all the data sources **110** discussed herein. **Table 2 gives a summary of the features with their type.** Each of the features shall be discussed below.

Table 2: Summary of features.

Feature	Type
NVD and DW description	TF-IDF on bag of words
CVSS	Numeric and Categorical
DW Language	Categorical
Presence of proof-of-concept	Binary
Vulnerability mention on ZDI	Binary
Vulnerability mention on DW	Binary

[0056] **NVD and DW description**. NVD description provides information on the vulnerability and what it allows hackers to do when they exploit it. DW description often provides rich context on what the discussion is about (mostly in forums rather than marketplaces since items are described in fewer words). Patterns can be learned based on this textual content. The description of published vulnerabilities was obtained from NVD between 2015 and 2016. The inventors also queried the DW database for CVE mentions between 2015 and 2016. This description was appended to the NVD description with the corresponding CVE. It was observed that some of the descriptions on DW are in a foreign language as discussed herein. The foreign text was translated to English using Google Translate API . The text features were vectorized using term frequency-inverse document frequency (TF-IDF) model, which creates a vocabulary of all the words in the description. The importance of a word feature increase the more times it occurs but is normalized by the total number of words in the description. This eliminates common words from being important features. The TF-IDF model was limited to

IPR2025-01398
CrowdStrike EX1002 Page 101

1000 most frequent words (there was no benefit in the performance by using more word features, it just adds to the computational cost).

[0057] **CVSS**. NVD provides a CVSS score and the CVSS vector from which the score is computed indicating the severity of each of the disclosed vulnerability. CVSS version 2.0 was used as a feature for a classifier (numeric type) implemented with the described exploitation prediction model. The CVSS vector lists the components from which the score is computed. The components of the vector include Access Complexity, Authentication, Confidentiality, Integrity and Availability. Access complexity indicates how difficult is to exploit the vulnerability once the attacker has gained access. It is defined in terms of three levels: High, Medium and Low. Authentication indicates whether authentication is required by the attacker to exploit the vulnerability. It is a binary identifier taking the values Required and Not Required. Confidentiality, Integrity and Availability indicate what loss the system would incur if the vulnerability is exploited. It takes the values None, Partial and Complete. All the CVSS vector features are categorical. The inventors vectorized these features by building a vector with all possible categories. Then if that category is present we insert "1" otherwise "0".

[0058] **Language**. DW feeds are posted in different languages. Four languages were found that are used in DW posts referencing vulnerabilities. These languages are English, Chinese, Russian, and Swedish. Since the number of samples was limited from every language, the text translation was used as described. However, translation can result in a loss of important information, but the impact of knowing the language is retained by using it as feature. The analysis on the languages of DW fees are shown and their variation in the exploitation rate.

[0059] **Presence of proof-of-concept**. The presence of proof-of-concept exploits in EDB increases the likelihood of a vulnerability being exploited. It was treated as a binary feature indicating whether proof-of-concept is present for a vulnerability or not.

[0060] **Vulnerability mention on ZDI**. Vulnerability mention on ZDI also increases the likelihood of it being exploited. Similar to proof-of-concept exploit a binary feature was used to denote whether a vulnerability was mentioned (disclosed) in ZDI before it is exploited. Vulnerability mention on DW. Vulnerability mention on DW also increases the likelihood of it being exploited. Binary feature indicating vulnerability mention on DW is considered as a feature.

IPR2025-01398
CrowdStrike EX1002 Page 102

[0061] On the other hand, a subset of the mentioned features comprising NVD and DW description, CVSS base score, and the language of DW posts are also used in embodiment B. In addition to those features, two more sets of features may be added. Those features include: (1) Common Weakness Enumeration (CWE), and (2) a set of features computed from the social network structure of users posting in a forum. A summary of each of these two sets of features is provided below.

[0062] **CWE**. It is a community-effort project comprising enumerating common software security weaknesses (categorical). These are categories of flaws that can be made during the software development and can exist in the software architecture, design, or code.

[0063] **DW Social Network Features**. This set of features contains measures computed from the social connections of users posting hacking-related content. The basic social network features (e.g., in-degree, out-degree) indicates how active a user is in the social graph. More advanced features measure the centrality of users in the social graph. Highly central users are more important; thus the vulnerability mentions should take more consideration. The features were computed for the set of users who explicitly mentioned one or more CVE-IDs in their posts.

VULNERABILITY ANALYSIS

[0064] For vulnerabilities used for evaluating embodiment A, the likelihood that a given vulnerability mentioned in a data source **110** will be exploited is analyzed. Time-based analysis is then provided based on the dates the exploited CVE's were mentioned in the data sources, and the *exploitation date* found in Symantec data. This analysis shows us how many of the vulnerability mentions online preceded their *exploitation date*. As mentioned herein, the vulnerabilities that were reported by Symantec without the exploitation date are 112; while those reported with their exploitation dates are 194. All exploited vulnerabilities are used as ground truth for the first analysis (likelihood of exploitation). However, in the time-based analysis the exploited vulnerabilities without reported dates were ignored since it cannot make any assumptions if they were detected before or after the vulnerability disclosed. Furthermore, we analyze our ground truth and compare it with other sources to identify the highly vulnerable software and system vendors.

IPR2025-01398
CrowdStrike EX1002 Page 103

Symantec reports attack signatures for vulnerabilities existing in certain products. The distribution of the affected vendors in every data source was studied to show the variation in vendor coverage attained from our data sources. This analysis was based on the vendor mentions by CPE data from NVD. A vulnerability can exist in variant software versions that run on different operating systems. Finally, a language-based analysis on the DW data was provided to shed light on some sociocultural factors present in darkweb and appear to have implications on the exploitability likelihood.

[0065] **Likelihood of Exploitation**. Before the likelihood of vulnerability mentions online is computed in the data sources **110** are exploited, it is important to understand the variability between these data sources **110** in covering exploited vulnerabilities. Higher variability between data sources **110** implies higher information gain and less redundancy. On the other hand, if an online mention of a vulnerability in a data source **110** does not increase the likelihood of being exploited as compared to NVD, the inclusion of that data source **110** adds no value to the prediction model. Table **3** shows the vulnerability exploitation probabilities for the vulnerabilities mentioned in a given data source **110**. This analysis emphasizes the value of *overt* data sources **110** in supplementing NVD data. As mentioned herein, around 2.4% of the vulnerabilities disclosed in NVD are exploited in the wild. Hence, including other sources can increase the likelihood of correctly predicting the vulnerability that will be exploited.

[0066] **Time-Based Analysis**. Most software systems receive repeated attacks to vulnerabilities known in the system long time after an exploit has been detected in the wild. According to a recent empirical study on exploit data recorded by Symantec, more than half of the attacked software and systems received the same attacks repeatedly over more than two years. Vulnerability may take long time between the date they are disclosed to the date patch is deployed. To develop sound metrics that assesses the vulnerability importance i.e. detects vulnerabilities before they are exploited in real world attack, vulnerability mentions are considered that appear before the exploitation date, or a short-time after the exploitation date such that the vulnerability is still not patched by many system administrators. Here, only the population of exploited vulnerabilities that are reported are analyzed with their *exploitation date* (n = 194).

IPR2025-01398
CrowdStrike EX1002 Page 104

[0067] FIG. **3** shows that more than 90% of cases, vulnerabilities are disclosed by NIST before any real-world attack is detected. In the other few cases, the attacks were launched in the wild before NIST published the vulnerabilities. This phenomena is attributed to three reasons: 1) the vulnerability information is sometimes leaked before the disclosure, 2) by the time NIST disclosed a vulnerability in NVD, other sources have already validated and disclosed it, then exploits are rapidly used it in real-world attacks, or 3) the attacker knew that what they were doing was successful and continued to exploit their targets until discovered. Additionally, ZDI and NVD have limited variation on the vulnerability disclosure dates (median is 0 days). It is important to note that as ZDI disclosures come from industry, that reserved CVE numbers are shown earlier here than in other sources.

Table 3: Number of vulnerabilities exploited, the number of exploited vulnerabilities, the fraction of exploited vulnerabilities that appeared in each source, and the fraction of total vulnerabilities that appeared in each source. Results are reported for vulnerabilities and exploited vulnerabilities appeared in EDB, ZDI, DW (distinct CVEs), CVEs in ZDI or DW, and results for intersection of the three sources.

	EDB	ZDI	DW	ZDI v DW	EDB v ZDI v DW
Number of vulnerabilities	799	824	378	1180	1791
Number of exploited vulnerabilities	74	95	52	140	164
Percentage of exploited vulnerabilities	21%	31%	17%	46%	54%
Percentage of total vulnerabilities	6.3%	6.5%	3.0%	9.3%	14.2%

[0068] In case of EDB database, almost all of the exploited vulnerabilities (with *exploitation date)* from the vulnerabilities that have proof-of-concept exploits archived in EDB were found in the wild within the first 100 days of proof-of-concept availability. Such short time period between proof-of-concept availability and actual attack in the real-world indicates that having a template for vulnerability exploitation (in this case proof-of-concept) makes it easier to be deployed in real world attacks. FIG. **4** shows difference in days between the availability of proof-of-concepts exploits and exploitation date. In case of DW database, more than 60% of the first-time mentions to the exploited CVE's are within 100 days before or 100 after the *exploitation dates*. The remaining mentions are within the 18 month time frame after the vulnerability *exploitation date* (see FIG. **5**).

IPR2025-01398
CrowdStrike EX1002 Page 105

[0069] **Vendor-Based Analysis**. It is important to note that Symantec reports vulnerabilities that attack the systems and software configurations used by their customers. In the time period studied, more than 84% and 36% of the exploited vulnerabilities reported by Symantec exist in products solely from, or run on, Microsoft and Adobe's products, respectively; whereas less than 16% and 8% of vulnerabilities published in NVD are related to Microsoft and Adobe, receptively. FIG. **6** shows the percentage from the exploited vulnerabilities that can affect each of the top five vendors in every data source. It is important to note that a vulnerability may affect more than a vendor (e.g., CVE-2016-4272 exists in Adobe Flash Player, and it allows attackers to execute arbitrary code via unspecified vectors and can affect products from all five vendors). Additionally, the absence of vulnerabilities detected in other important systems and software vendors from Symantec's dataset does not imply that they have not been exploited; rather, they are not detected by Symantec (false positives). Furthermore, the presence of some operating systems vendors (e.g., Linux) in the exploited vulnerabilities does not necessarily imply good coverage of Symantec's data to these systems; however, other exploited products can run on these operating systems.

[0070] DW data appears to have more uniform vendor coverage. Only 30% and 6.2% of the vulnerabilities mentioned in DW during the time period are related to Microsoft and Adobe, receptively. Additionally, ZDI favors products from these two vendors (Microsoft is 57.8%, and Adobe 35.2%), as well as Apple's products (40.6%). This provides evidence that each data source covering vulnerabilities targeting varying sets of software vendors.

[0071] **Language-Based Analysis**. The users posting in DW hacking forums, and the vendors selling in marketplaces, are found to be discussing technical details about disclosed vulnerabilities, and trading real-world exploits. Interestingly, it was found that notable variations on the exploitation likelihood conditioned on the language used in DW data feeds are referencing CVE's. In DW feeds, four languages are detected with different vulnerability posts and items distributions. FIG. **7** shows the percentage of vulnerability mentions and the fraction of exploited vulnerabilities given it is mentioned by each language. Expectedly, English and Chinese, have the highest numbers of vulnerabilities (n = 242, and *n* = 112, respectively). However, vulnerability posts in Chinese exhibit the lowest exploitation rate (about 10%) followed by English (about 13%). The other two

IPR2025-01398
CrowdStrike EX1002 Page 106

languages are Russian and Swedish with few vulnerability mentions, but both exhibit very high exploitation rates. About 46% of the vulnerabilities mentioned in Russian where exploited (n = 6), and about 19% for vulnerabilities mentioned in Swedish (n = 2). Even though languages imply the likelihood of exploitation, they can be viewed as a double-edged sword since the language-dependent sets of textual features need considerable size of training corpus, and DW data feeds are insufficient for training classifiers for each language. As mentioned herein, the translation of the DW text is used after preprocessing (special character removal, stemming and bag of word) as well as the vulnerability description from NVD to train and test our model. This helps in capturing valuable textual features related to hacking and exploiting vulnerabilities, but other cultural-dependent features may go uncaptured. To account for any lose from the limitation of training samples, we use the DW mention languages as features.

EXPERIMENTAL SETUP (embodiment A)

[0072] Three experiments were performed with the models to determine effectiveness in prediction of exploitability. First, the model was compared to a benchmark done under conditions very close to work previously presented. For present model, it was found that Random Forest (RF) gives us the best F1 measure. Random forest is an ensemble method proposed by Breiman. It is based on the idea of generating multiple predictors (decision trees in this case) which are then used in combination to classify a new disclosed vulnerability. The strength of the random forest lies in introducing randomness to build each classifier and using random low dimensional subspaces to classify the data at each node in a classifier. A random forest was used that combines bagging for each tree with random feature selection at each node to split the data. The final result is therefore an ensemble of decision trees each having their own independent opinion on class labels (exploited or not exploited) for a given disclosed vulnerability. A new vulnerability is classified independently by each tree and assigned a class label best fit for it. Multiple decision trees may result in having multiple class labels for the same data sample; hence, to take the final decision a majority vote is taken and the class label with most votes is assigned to the vulnerability.

IPR2025-01398
CrowdStrike EX1002 Page 107

Performance Evaluation (embodiment A)

[0073] The classifiers were evaluated based on two classes of metrics that have been used in previous work. The first class is used to demonstrate the performance achieved on the minority class (in our case here 1.2%). The metrics under this class are precision and recall. They are computed as reported in Table **4**. Precision is defined as the fraction of vulnerabilities that were exploited from all vulnerabilities predicted to be exploited by the present model. It highlights the effect of mistakenly flagging non-exploited vulnerabilities. Recall is defined as the fraction of correctly predicted exploited vulnerabilities from the total number of exploited vulnerabilities. It highlights the effect of unflagging important vulnerabilities that were used later in attacks. For highly imbalanced data, these metrics give an intuition regarding how well the classifier performed on the minority class (exploited vulnerabilities). The F1 measure is the harmonic mean of precision and recall. It summarizes the precision and recall in a common metric. The F1 measure can be varied based on the trade-off between precision and recall. This trade-off is dependent on the priority of the applications. If keeping the number of incorrectly flagged vulnerabilities to a minimum is a priority, then high precision is desired. To keep the number of undetected vulnerabilities that are later exploited minimum, high recall is desired. We further report Receiver Operating Characteristics (ROC) curve as well as Area Under Curve (AUC) of the classifier, which is close to perfect (95%). ROC graphically illustrates the performance of our classifier by plotting the true positive rate (TPR) against the false positive rate (FPR) at various thresholds of the confidence scores the classifier outputs. In binary classification problems, the overall TPR is always equivalent to recall for the positive class while FPR is the number of *not exploited* vulnerabilities that are incorrectly classified *as exploited* from *all not exploited* samples. ROC is a curve; thus, AUC is the area under ROC. The higher, the better (i.e., a classifier with AUC =1 is a perfect classifier).

IPR2025-01398
CrowdStrike EX1002 Page 108

Table 4: Evaluation metrics. TP - true positives, FP - false positives, FN - false negatives, TN - true negative.

Metric	Formula
Precision	$\frac{TP}{TP+FP}$
TPR (recall in case of binary classification)	$\frac{TP}{TP+FN}$
F1	$2*\frac{precision*recall}{precision+recall}$
FPR	$\frac{FP}{FP+TN}$

Results (embodiment A)

[0074] **Benchmark Test**. The present model was compared to recent works on using vulnerabilities mentioned on twitter to predict if such vulnerabilities would be exploited or not. They use Support Vector machine (SVM) as their classifier, the present model works best with Random Forest classifier. Although it would be straightforward to think that our approach would achieve better performance than previous work, this work is only compared because: (1) to the best of our knowledge, there is no existing work on predicting exploits in the wild using DW data, and (2) it is compared with all major approaches, and using feeds from social media currently is the best one. The inventors restrict the training and evaluation of their classifier to vulnerabilities targeting Microsoft and Adobe products as Symantec does not have attack signatures for all the targeted platforms. They perform a 10-fold stratified cross validation, where the data is partitioned into 10 parts while maintaining the class ratio in each part, they train on 9 parts and test on the remaining one. The experiment is repeated for all 10 parts. Hence, each sample gets tested at least once.

[0075] For comparison, the same experiment is performed, under highly similar assumptions. Exploited vulnerabilities are all used regardless of whether the date is reported by Symantec or not. In the present case, 2056 vulnerabilities target Microsoft and Adobe products. Out of 2056 vulnerabilities, 261 are exploited, fraction that is consistent with previous work. A 10-fold stratified cross-validation is performed. The precision-recall curve is plotted for the present model (see FIG. **8**). The precision-recall curve shows us the trade-off between

IPR2025-01398
CrowdStrike EX1002 Page 109

precision and recall for different decision threshold. Since F1 measure is not reported, the precision-recall curve reported is used to draw comparison. By maintaining the recall value constant, how the precision varies are compared. Table **5** shows the comparison between the two models by comparing precision for different values of recall. For a threshold of 0.5 we get an F1 measure of 0.44 with precision 0.53 and recall 0.3. Maintaining the recall, the precision displayed in the graph is 0.3 significantly lower than 0.4. The same experiment is performed on different recall values to compare the precision. At each point, higher precision is obtained than the previous approach.

Table 5: Precision comparison between * and proposed model while keeping the recall constant.
* Numbers derived from Figure 6.a. from [44]

Metric	Precision [44]*	Precision (This paper)
Recall = 0.20	0.30	0.41
Recall = 0.40	0.18	0.40
Recall = 0.70	0.10	0.29

[0076] **Avoiding temporal intermixing**. It should be understood that the temporal sequencing of exploit warning could lead to future events being used to predict past ones. This could lead to inaccurate results, especially in the case where there is strong temporal dependencies in the data - which argues is the case for social media. In this experiment we sort the vulnerabilities according to their disclosed dates on NVD. The first 70% was reserved for training and the rest for testing.

[0077] For a baseline comparison we use the CVSS version 2.0 score to classify whether a vulnerability will be exploited or not based on the severity score assigned to it. The CVSS score has been used as a baseline in previous studies. CVSS tends to classify on the side of caution i.e. it tends to predict more vulnerabilities that will be exploited resulting in false positives. FIG. **9** shows the precision-recall curve for the CVSS score. It is computed by varying the decision threshold that decides whether to predict a vulnerability as exploited or not. CVSS gives high recall with very low precision which is not desired for real-world exploit

IPR2025-01398
CrowdStrike EX1002 Page 110

detection. The best F1 measure that could be obtained was 0.15. The experiment is now performed using our proposed model.

[0078] FIG. **10** shows the performance comparison between the proposed model using the random forest classifier and the CVSS score. The best F1 measure for the CVSS score is 0.15. The model outperforms the CVSS score with an F1 measure of 0.4 with precision 0.45 and recall 0.35.

[0079] The performance on the minority class is promising when realizing that the class imbalance is very severe and the ground truth is not perfect. The random forest classifier is able to score an average of F1 of 0.89 for 10 runs of randomly under-sampled training and testing datasets. However, this is not considered to reflect the actual performance for real-world deployment. Additionally, our classifier show very high TPR (90%) at low FPR (13%) and AUC of 95% as depicted in FIG. **11**. This is comparable to previous work in predicting cyber threats, though in different prediction tasks, (TPR: 90%, and FPR: 10%), results reported in [49](TPR: 66%, and FPR: 17%), as well as reported results in [31](TPR: 70%, and FPR: 35%).

[0080] **Evaluation with Individual Data Sources**. What this effect does was studied introducing each data source has on the prediction of vulnerabilities mentioned in that source. This is an important to see if adding a particular data source benefits the vulnerabilities that have been mentioned in that data source. It was found that time based split used in the previous experiment leaves very few vulnerabilities mentioned in these data sources in the test set (ZDI: 18, DW: 4, EDB: 2). Hence the increase the numbers by (1) performing a 10-fold cross validation without sorting the vulnerabilities (2) the ground truth was increased by considering the exploited vulnerabilities that did not have exploit date (these were removed from earlier experiments since it was not clear whether these were exploited before or after the vulnerability was exploited). Using these two techniques, we have 84 vulnerabilities mentioned in ZDI that have been exploited, 57 in EDB, and 32 in DW. The results (precision, recall and F1) were reported for the vulnerabilities mentioned in each data source. Also, the prediction of these vulnerabilities was mentioned by using only NVD features. For the vulnerabilities mentioned in DW, DW features were only considered along with NVD, The model predicts 12 vulnerabilities as exploited with a precision of 0.67 and recall of 0.375. By only considering the NVD features, predicts 12 vulnerabilities as exploited with a

IPR2025-01398
CrowdStrike EX1002 Page 111

precision of 0.23 and recall of 0.38. Hence using the DW features, the precision improved significantly from 0.23 to 0.67. Table **6** shows the precision-recall with corresponding F1 measure. DW information was thus able to correctly identify the positive sample mentioned in DW with higher precision.

Table 6: Precision, Recall, F1 measure for vulnerabilities mentioned on DW, ZDI, and EDB.

Source	Case	Precision	Recall	F1 measure
DW	NVD	0.23	0.38	0.27
	NVD + DW	0.67	0.375	0.48
ZDI	NVD	0.16	0.54	0.25
	NVD + ZDI	0.49	0.24	0.32
EDB	NVD	0.15	0.56	0.24
	NVD + EDB	0.31	0.40	0.35

[0081] For ZDI, 84 vulnerabilities were mentioned in it. By just utilizing NVD features, we get an F1 measure of 0.25 (precision: 0.16, recall: 0.54) as compared to addition of ZDI feature with F1 measure of 0.32 (precision: 0.49, recall: 0.24), a significant improvement in precision. Table **6** also shows the precision-recall with corresponding F1 measure for samples mentioned on ZDI,

[0082] A similar analysis was performed for the vulnerabilities that have proof-of-concept available on EDB. For EDB we have 57 vulnerabilities with proof-of-concept. By just utilizing NVD features an F1 measure of 0.24 (precision: 0.15, recall: 0.56) was obtained as compared to addition of EDB feature with F1 measure of 0.35 (precision: 0.31, recall: 0.40), a significant improvement in precision as shown in Table **6**.

[0083] **Feature Importance**. To better explain the choices to the features that were examined and provide an understanding on where the prediction power most attributed to, the features that have the most contribution to the prediction performance were reported. A feature vector for a sample has 28 features computed from the non-textual data (summarized in Table **2**) as well as the textual features - TF-IDF computed from the bag of words for the 1000 words that have the highest frequency in NVD description and DW. For each of the features, the Mutual

IPR2025-01398
CrowdStrike EX1002 Page 112

Information (MI) was computed, which computes how much a variable (here a feature x_i) tells about another variable (here the class label $y \in$ *{exploited, not exploited}*). The features that contribute the most from the non-textual data are *{language_Russian = true, has_DW = true, has_PoC =false}*. In addition, the features that contribute the most from the textual data are the words *{demonstrate, launch, network, xen, zdi, binary, attempt}*. All of these features received MI scores over 0.02.

ADVERSARIAL DATA MANIPULATION (embodiment A)

[0084] The effects of adversarial data manipulation on DW data were studied. For EDB, we only consider vulnerabilities that have verified proof-of-concepts. ZDI also publishes only verified vulnerabilities; hence there is a very small chance of manipulating these data sources. On the other hand, the public nature of DW marketplaces and forums gives an adversary the ability to poison the data used by the classifier. They can achieve it by adding vulnerability discussions on these platforms with the intent of deceiving the classifier to predict exploitation by that vulnerability and having high false positives. Previous work discusses how an adversary can influence a classifier by manipulating the training data.

[0085] In the present prediction model, we use the presence of the vulnerability in darkweb, language of the market/forum it was mentioned and the vulnerability description as features. An adversary could easily post discussions regarding vulnerabilities he does not intent or has information that they will not be exploited. To study the influence of such noise on the performance of the model was conducted in two strategies:

[0086] 1. **Adversary adding random vulnerability discussion**. In this strategy, the adversary initiates random vulnerability discussions on DW and reposts them with a different CVE number. So the CVE mentions on DW increases. For the present experiment, two cases were considered with different amounts of noise added. In case (1), it is assumed that the noise is present in both training and testing data. Varying fractions of noise (5%, 10%, 20% of the total data samples) randomly distributed in training and testing data were considered. The experimental setup follows conditions discussed herein. Vulnerabilities are first sorted according to time, and the first 70% are reserved for training and the remaining for testing. FIG. **12A** shows the ROC curve showing the false positive rate (FPR) vs the true

IPR2025-01398
CrowdStrike EX1002 Page 113

positive rate (TPR). For different amount of noise introduced, the present model still maintains a high TPR with low FPR and AUC ≥ 0.94, a performance similar to the experiment without adding noise. This shows that the model is highly robust against noise such that it learns good representation of the noise in the training dataset then can distinguish them in the testing dataset.

[0087] For case (2), we randomly add vulnerability discussion found on DW with different CVE number to only the test data and repeat the same experiment. FIG. **12A** shows the ROC plot. In the case, even though there is a slight increase in the FPR, the performance is still on par with the experiment without noise (AUC 0.87). Hence, noisy sample affect the prediction model slightly, if no noise was introduced in the training data.

[0088] **2. Adversary adding vulnerability discussion similar to NVD**. In the previous strategy, the adversary was adding vulnerability discussions randomly without taking into account the actual capability of the vulnerability. For instance, CVE-2016-3350 is a vulnerability in Microsoft Edge as reported by NVD. If the vulnerability is mentioned on DW as noise by an adversary but targeting Google Chrome, then it might be *easy* for the prediction model to detect it as seen in previous experiments. But, what if the adversary crafts the vulnerability discussion such that it is a copy of the NVD description or consistent with the NVD description? In this strategy, the adversary posts the NVD description with the CVE number on DW. For case (1), this noise is considered to be randomly distributed in both training and testing. Figure 12(b) shows the ROC curves for different levels of noise. The performance decreases as the number of noisy samples increases, but there is no significant decline (AUC 0.88).

[0089] The experiment was repeated with adding noise only in the test data for case (2). In this experiment the most performance was observed with AUC of 0.78 for 20% noise (see FIG. **12B**). This shows that adding correct vulnerability discussions does affect the prediction model but with large number of such samples. Also, effect can be countered by adding such noisy samples in the training data as well for the model to learn from.

[0090] An important point to consider is an adversary would need to add a large number of noisy samples to drive down the performance of the prediction model. Previous research on using data feeds like Twitter for exploit prediction mention that an adversary can purchase large number of twitter accounts and flood

IPR2025-01398
CrowdStrike EX1002 Page 114

twitter with vulnerability mentions. In DW markets and forums, creation of accounts needs verification and in some cases demonstration of skills to get in. While fake accounts are often themselves sold on the darkweb, it is difficult to purchase and maintain thousands of such fake accounts to post in them. Also, if one person is posting a large volume of discussions with CVE mentions, he/she can be identified from their *username* or can be removed from the market/forum if many of their posts get down voted as being not relevant. It is also important to note that such forums also function as a meritocracy — where users who contribute more are held in higher regard (which also makes it difficult to flood discussions with such information).

DISCUSSION (embodiment A)

[0091] Viability of the Model and Cost of Misclassification The performance achieved by the model as a first-line defense layer is very promising. Recall that random forest classifier outputs a confidence score for every testing sample. A threshold can be set to identify the decision boundary. It should be noted that all the results reported in this disclosure are achieved based on hard-cut thresholds such that all samples that are assigned confidence score greater than a threshold *thr* are predicted as exploited. Blindly relying on a hard-cut threshold may not be a good practice in real-world exploits prediction; rather, *thr* should be varied in accordance to other variables within the organization such that different thresholds can be set to different vendors (i.e., thr_{ven1}, thr_{ven2}), or information systems (i.e., thr_{sys1}, thr_{sys2}). For instance, if an organization hosts an important website on an Apache server, and the availability of that site is of the top priority for that organization, then any vulnerability in Apache server should receive high attention and put forward to remediation plan regardless of other measures. Other vulnerabilities, tens of which are disclosed in daily bases, may exist in many other systems within the organization. Since it is very expensive to be responsive to that many security advisories (e.g., some patches may be unavailable, some systems may need to be taken offline to apply patches), exploitation assessment measures can help in quantifying the risk and planning mitigations. Risk is always thought of as a function of likelihood (exploitation) and impact. The cost of classifying negative samples as *exploited* is the effort made to have it fixed. This mostly involves patching, or other remediation such as controlling access or blocking network ports. Similarly, the cost of misclassification depends on the impact incurs. For example, if

IPR2025-01398
CrowdStrike EX1002 Page 115

two companies run the same database management system *s,* and one hosts a database with data about all business transactions for that company while the other host a database with data of that is of little value to the company, then the resulting cost of a data breach is significantly different.

[0092] **Model Success and Failure Cases**. By analyzing the false negatives and false positives, an understanding is gained as to why and where our model performs well as well as why and where it suffers. The 10 exploited vulnerabilities (about 18% of the *exploited* samples in the testing dataset) that received the lowest confidence scores seem to have common features. For example 9 of these 10 exist in products from Adobe, namely Flash Player (5 vulnerabilities) and Acrobat Reader (4 vulnerabilities). Flash Player's vulnerabilities look to have very similar description from NVD, and the same thing is observed for Acrobat Reader's product. It was also observed that they were assigned CVE-IDs at the same date (April 27th, 2016), and 7 out of these 9 were published at the same data as well (July 12th, 2016), and assigned a CVSS base score = 10 (except for one, which is 7). The other vulnerability exist in Windows Azure Active Directory (CVSS score = 4.3). Out of these 10 vulnerabilities, only one had a verified PoC archived on EDB before it was detected in the wild, and another one had a ZDI mention, while none was mentioned in DW. Misclassifying these vulnerabilities are attributed to the limited representation of these samples in the training data set. This observation signifies the importance of avoiding experiments on time-intermixed data, a point discussed herein.

[0093] The false positive samples were looked into that receive high confidence — samples where the model predicted *as exploited* while they are not. For the random forest classifier, with F1 = 0.4, all the false positives (about 40 out of about 3600 vulnerabilities) exist in products from Microsoft although the model does not use the vendor as feature. The model seems to be able to infer the vendor from other textual features. The inventors posit that this level of overfitting is unavoidable and marginal, and attribute this largely to the limitations on the ground truth. Though the model is highly generalizable, examples of vulnerabilities were found from other vendors with confidence scores close to *thr* that was set; however, it cannot be assumed that they are exploited.

IPR2025-01398
CrowdStrike EX1002 Page 116

Hacker Social Network (embodiment B)

[0094] Embodiment B of present disclosure, adopts the same assumption made in much of the current literature related to DW data, in which that the same usernames (case insensitive) across different DW sites is assumed to belong to the same person(s). This assumption allows for generating one network comprising a large number of DW sites as opposed to a social network for each site.

[0095] **Users' Social Graph**. Formally, the users' social graph $G=(V,E)$ is a weighted, directed graph with no self-loops (i.e., every edge has a weight; every edge points away from one node to another node; there exists at most one edge between any pair of nodes). V is the set of vertices (DW users) and E is the set of edges.

[0096] Then, an edge is created (with weight = 1) from user v_i to v_j and labeled with the date of v_i's posting date only if: (1) $v_i \neq v_j$, (2) both v_i and v_j have posts in I_X, and v_i has posted after v_j, (3) the number of posts between v_i's post and v_j's post in I_X is less than *thr* (it is set to be 10 in all experiments in this disclosure), and (4) there is no existing edge originating from v_i to v_j and labeled with the same date. Once the edges are created, they are added to a multi-directed graph with parallel edges of weights = 1. The multi-graph is then transformed to a directed graph G by summing the weights of the parallel edges pointing at the same direction.

Social Network Measures (embodiment B)

[0097] After creating the social network, measures are computed derived from the network structure. In this disclosure, three categories of social network measures are considered:

[0098] **Network Structure Measures**: the measures under this category are: (1) *In-degree*: the number of edges pointing to the user, (2) *Out-degree*: the number of edges originated from the user, (3) *Sum of In-degree weights*: the sum of the weights for all edges pointing to the user, (4) *Sum of out-degree weights*: the sum of the weights for all edges pointing away from the user. These measures describe the type of activities in which the user engages. For example, higher in-degree than out-degree may indicate the user tendency towards creating new topics or posting under topics short time after they are created.

IPR2025-01398
CrowdStrike EX1002 Page 117

[0099] **Centrality Measures:** three measures are computed: (1) *In-degree centrality*: it measures the popularity of a user v_i by normalizing v_i's in-degree by the maximum possible in-degree, (2) *Out-degree centrality*: measures how actively a user v_i replies to others by normalizing v_i's out-degree measure by the maximum possible out-degree, (3) *Betweenness centrality*: for a user v_i, Betweenness centrality measures the importance of v_i by computing the fraction of shortest paths between all pairs of users that pass through v_i.

[00100] **Importance Measures:** the number of connections user v_i has with other users, by itself, may not be indicative of importance; rather, v_i is important if his/her posts make other important users reply. Hence, influence metrics incorporate the centrality of users with outgoing edges to v_i into v_i's centrality (i.e., if an important user v_j replies to v_i, then the importance of v_i increases). Two measures are computed under this category: (1) *Eigenvector centrality*: measures the importance of v_i by assigning a centrality proportional to the sum of in-neighbors' centralities. Eigenvector centrality of v_i is the ith value of the eigenvector C_e corresponding to the largest eigenvalue of the network adjacency matrix A^t, and (2) *Pagerank centrality*: measures the centrality of v_i by incorporating fractions of the centralities of in-neighbors, such that each of v_i's in-neighbors passes the value of his/her centrality divided by the number of outgoing edges.

EXPERIMENTAL SETUP (embodiment B)

[00101] In the present embodiment, only the vulnerabilities mentioned on the considered DW forums are considered. Additionally, the same experimental setup as used in embodiment A is used to examine the present embodiment. However, two additional steps are taken here; (1) the vulnerabilities that were mentioned by users with no communication history are removed, and (2) in each of the experiments, reported an average performance of five runs. The resultant dataset contains 157 distinct vulnerabilities, 24 of which have the class label *exploited*. Additionally, embodiment B of the present disclosure is evaluated on the exact same performance metrics as in embodiment A.

IPR2025-01398
CrowdStrike EX1002 Page 118

Results (embodiment B)

[00102] **Experiments under Real-World Conditions**. In this set of experiments, the vulnerabilities were sorted by their DW date, then we train our classifiers on the vulnerabilities mentioned before June 2016 (125 vulnerabilities), and test on the vulnerabilities from June 2016 to March 2017 (32 vulnerabilities, only 3 are *exploited*. The classification performance achieved by our RF model has an average precision of 0.57, recall of 0.93, and F1 of 0.67. The same classifier is able to achieve on average AUC of 0.95. We note that the results of 5 runs show relatively high variance due to the small number of samples on which the models are tested. The lower score of precision is attributed to the fact that Symantec's data is biased towards reporting exploits targeting vulnerabilities that exist in software products from certain software vendors as discussed earlier. Since our model is found to predict vulnerabilities as being exploited from other vendors as well, we believe that some false positives were actually exploited in the wild but never detected by Symantec.

[00103] **Ablation Test and Cross-Validation**. Since the number of vulnerabilities in the testing dataset in the previous experiment is relatively small, stratified 5-fold cross-validation was applied on the whole dataset. In this experiment, the samples are intermixed; hence these conditions do not reflect the conditions of real-world streaming prediction (i.e., predicting the likelihood of exploitation at the time of the vulnerability mention). The average F1 achieved is 0.72, with a precision of 0.61, a recall of 0.89, and an AUC of 0.88.

[00104] To measure the impact of individual feature sets on the overall classification performance, two tests were applied: (1) an ablation test (depicted in FIG. **16**) where the change in precision, recall, F1, and AUC is recorded when each set of features is removed from the prediction model, and (2) a test on individual feature sets (depicted in FIG. **17**) where the classification performance is reported for models trained on only one set features at a time. In the ablation test, when the set of social network features were removed from the model, some decrease in performance was recorded as depicted in FIG. **16**. In the individual feature tests, the social network measures resulted in an improvement in performance that is significantly higher than the improvement recorded from the inclusion of other feature sets. It was noted that the simple classifier, which labels all vulnerabilities as being

IPR2025-01398
CrowdStrike EX1002 Page 119

exploited, results in a precision of 0.16, a recall of 1, at an F1 of 0.27 and an AUC of 0.5.

CONCLUSION

[00105] A study of aggregating early signs from online vulnerability mentions was conducted for predicting whether a particular disclosed vulnerability will be exploited. The present prediction model is robust and useful as compared against previously introduced models that combine information from social media sites like twitter for exploit prediction. The results show that while maintaining recall significantly high precision was achieved in predicting exploits.

PERFORMANCE

[00106] The performance of several standard supervised machine learning approaches was compared for exploit prediction models examined in the first embodiment. Parameters for all approaches were set in a manner to provide the best performance. The scikit-learn Python package was used. Other approaches for comparison will now be discussed.

Supervised Learning Approaches

[00107] **Support Vector Machine (SVM)**. Support vector machine (SVM) was proposed by Vapnik . SVM works by finding a separating margin that maximizes the geometric distance between classes (in the instant case, exploited and not exploited). The separating margin is termed as hyperplane. When a separating plane cannot be found to distinguish between the two classes, the SVM cost function includes a regularization penalty and a slack variable for the misclassified samples. Varying these parameters, trade-off between precision and recall can be observed.

[00108] **Naive Bayes Classifier (NB)**. Naive Bayes is a probabilistic classifier which uses Bayes theorem with independent attribute assumption. During training we compute the conditional probabilities of a sample of a given class having a certain attribute. The prior probabilities for each class i.e. fraction of the training data belonging to each class was computed. Naive Bayes assumes that the attributes are statistically independent hence the likelihood for a sample *S*

IPR2025-01398
CrowdStrike EX1002 Page 120

represented with a set of attributes a associated with a class c is given as,

$$\Pr(c|S) = P(c) \times \prod_{i=1}^{d} \Pr(a_i|c)$$

[00109] **Bayesian Network (BN)**. BN is a generalized form of NB such that not all features are assumed to be independent. Rather, variable dependencies are modeled in a graph leaned from the training data.

[00110] **Decision Tree (DT)**. Decision tree is a hierarchical recursive partitioning algorithm. We build the decision tree by finding the best split attribute i.e. the attribute that maximizes the information gain at each split of a node. In order to avoid over-fitting, the terminating criteria is set to less than 5% of total samples.

[00111] **Logistic Regression (LOG-REG)**. Logistic regression classifies samples by computing the odds ratio. The odds ratio gives the strength of association between the attributes and the class.

Results

[00112] The temporal information for all the classifiers was maintained. The disclosed vulnerabilities were sorted according to the time they were posted on NVD. The first 70% was reserved for training and the remaining for testing.

Table 7: Precision, Recall, F1 measure for RF, SVM, LOG-REG, DT and NB for predicting whether a vulnerability will be exploited or not.

Classifier	Precision	Recall	F1 measure
RF	0.45	0.35	0.40
BN	0.31	0.38	0.34
SVM	0.28	0.42	0.34
LOG-REG	0.28	0.4	0.33
DT	0.25	0.24	0.25
NB	0.17	0.76	0.27

[00113] Table **7** shows the comparison between the classifiers with respect to precision, recall and F1 measure. Random forest (RF) performs the best with F1 measure of 0.4 as compared to SVM: 0.34, BN: 0.34, LOG-REG: 0.33, DT:

IPR2025-01398
CrowdStrike EX1002 Page 121

0.25, and NB: 0.27. An interesting point to note, even though RF has the best F1 measure it does not have the best recall, NB does. RF was chosen as having high precision makes the model reliable as compared to low precision which results in a lot of false positives. The approach going forward that might improve recall without significantly affecting precision is to use an ensemble of different classifiers, where one approach helps the other approach to make the right decision. This way the strength of different classifiers can be combined and avoid their weaknesses.

Addressing Class Imbalance

[00114] The problem of class imbalance has gained lot researchers' attention because it exists in many real-world applications. This has resulted in large number of proposed approaches. Since our dataset is highly imbalanced (the percentage of minority class is only 1.2%), SMOTE is used to measure the improvement in classification performance. SMOTE over-samples the exploited vulnerabilities by creating synthetic samples with features similar to the exploited vulnerabilities. This data manipulating is only applied to the training set. Applying SMOTE, no performance improvement is achieved for our RF classifier. However, SMOTE introduces a considerable improvement with Bayesian Network (BN) classifier. Table **8** reports different over-sampling sizes and the change in performance. Over exaggerating in creating synthetic positive samples can overpower the real samples leading to reduction in the model performance.

Table 8: Performance improvement attained by applying SMOTE for BN classifier using different over-sampling percentages for the exploited samples

Over-Sampling Percentage	Precision	Recall	F1 measure
100%	0.37	0.42	0.39
200%	0.40	0.44	0.42
300%	0.41	0.40	0.40
400%	0.31	0.40	0.35

IPR2025-01398
CrowdStrike EX1002 Page 122

[00115] FIG. **18** is an example schematic diagram of a computing device **700** that may implement various methodologies and functions discussed herein. For example, the computing device **700** may be configured by the application **102** to implement any embodiments of the described exploit prediction model. The computing device **700** includes a bus **701** (i.e., interconnect), at least one processor **702** or other computing element, at least one communication port **703**, a main memory **704**, a removable storage media **705**, a read-only memory **706**, and a mass storage device **707**. Processor(s) **702** can be any known processor, such as, but not limited to, an Intel® Itanium® or Itanium 2® processor(s), AMD® Opteron® or Athlon MP® processor(s), or Motorola® lines of processors. Communication port **703** can be any of an RS-232 port for use with a modem based dial-up connection, a 10/100 Ethernet port, a Gigabit port using copper or fiber, or a USB port. Communication port(s) **703** may be chosen depending on a network such as a Local Area Network (LAN), a Wide Area Network (WAN), or any network to which the computer device **700** connects. Computing device may further include a transport and/or transit network **755,** a display screen **760**, an I/O port **740**, and an input device **745** such as a mouse or keyboard.

[00116] Main memory **704** can be Random Access Memory (RAM) or any other dynamic storage device(s) commonly known in the art. Read-only memory **706** can be any static storage device(s) such as Programmable Read-Only Memory (PROM) chips for storing static information such as instructions for processor **702**. Mass storage device **707** can be used to store information and instructions. For example, hard disks such as the Adaptec® family of Small Computer Serial Interface (SCSI) drives, an optical disc, an array of disks such as Redundant Array of Independent Disks (RAID), such as the Adaptec® family of RAID drives, or any other mass storage devices, may be used.

[00117] Bus **701** communicatively couples processor(s) **702** with the other memory, storage, and communications blocks. Bus **701** can be a PCI / PCI-X, SCSI, or Universal Serial Bus (USB) based system bus (or other) depending on the storage devices used. Removable storage media **705** can be any kind of external hard drives, thumb drives, Compact Disc – Read Only Memory (CD-ROM), Compact Disc – Re-Writable (CD-RW), Digital Video Disk – Read Only Memory (DVD-ROM), etc.

IPR2025-01398
CrowdStrike EX1002 Page 123

[00118] Embodiments herein may be provided as a computer program product, which may include a machine-readable medium having stored thereon instructions which may be used to program a computer (or other electronic devices) to perform a process. The machine-readable medium may include, but is not limited to optical discs, CD-ROMs, magneto-optical disks, ROMs, RAMs, erasable programmable read-only memories (EPROMs), electrically erasable programmable read-only memories (EEPROMs), magnetic or optical cards, flash memory, or other type of media/machine-readable medium suitable for storing electronic instructions. Moreover, embodiments herein may also be downloaded as a computer program product, wherein the program may be transferred from a remote computer to a requesting computer by way of data signals embodied in a carrier wave or other propagation medium via a communication link (e.g., modem or network connection).

[00119] As shown, main memory **704** may be encoded with the application **102** that supports functionality discussed above. In other words, aspects of the application **102** (and/or other resources as described herein) can be embodied as software code such as data and/or logic instructions (e.g., code stored in the memory or on another computer readable medium such as a disk) that supports functionality according to different embodiments described herein. During operation of one embodiment, processor(s) **702** accesses main memory **704** via the use of bus **701** in order to launch, run, execute, interpret, or otherwise perform processes, such as through logic instructions, executing on the processor **702** and based on the application **102** stored in main memory or otherwise tangibly stored.

[00120] The description above includes example systems, methods, techniques, instruction sequences, and/or computer program products that embody techniques of the present disclosure. However, it is understood that the described disclosure may be practiced without these specific details. In the present disclosure, the methods disclosed may be implemented as sets of instructions or software readable by a device. Further, it is understood that the specific order or hierarchy of steps in the methods disclosed are instances of example approaches. Based upon design preferences, it is understood that the specific order or hierarchy of steps in the method can be rearranged while remaining within the disclosed subject matter. The accompanying method claims present elements of the various steps in a sample order, and are not necessarily meant to be limited to the specific order or hierarchy presented.

IPR2025-01398
CrowdStrike EX1002 Page 124

[00121] The described disclosure may be provided as a computer program product, or software, that may include a machine-readable medium having stored thereon instructions, which may be used to program a computer system (or other electronic devices) to perform a process according to the present disclosure. A machine-readable medium includes any mechanism for storing information in a form (e.g., software, processing application) readable by a machine (e.g., a computer). The machine-readable medium may include, but is not limited to optical storage medium (e.g., CD-ROM); magneto-optical storage medium, read only memory (ROM); random access memory (RAM); erasable programmable memory (e.g., EPROM and EEPROM); flash memory; or other types of medium suitable for storing electronic instructions.

[00122] Certain embodiments are described herein as including one or more modules. Such modules are hardware-implemented, and thus include at least one tangible unit capable of performing certain operations and may be configured or arranged in a certain manner. For example, a hardware-implemented module may comprise dedicated circuitry that is permanently configured (e.g., as a special-purpose processor, such as a field-programmable gate array (FPGA) or an application-specific integrated circuit (ASIC)) to perform certain operations. A hardware-implemented module may also comprise programmable circuitry (e.g., as encompassed within a general-purpose processor or other programmable processor) that is temporarily configured by software or firmware to perform certain operations. In some example embodiments, one or more computer systems (e.g., a standalone system, a client and/or server computer system, or a peer-to-peer computer system) or one or more processors may be configured by software (e.g., an application or application portion) as a hardware-implemented module that operates to perform certain operations as described herein.

[00123] Accordingly, the term "hardware-implemented module" or "module" encompasses a tangible entity, be that an entity that is physically constructed, permanently configured (e.g., hardwired), or temporarily configured (e.g., programmed) to operate in a certain manner and/or to perform certain operations described herein. Considering embodiments in which hardware-implemented modules are temporarily configured (e.g., programmed), each of the hardware-implemented modules need not be configured or instantiated at any one instance in time. For example, where the hardware-implemented modules comprise

IPR2025-01398
CrowdStrike EX1002 Page 125

a general-purpose processor configured using software, the general-purpose processor may be configured as respective different hardware-implemented modules at different times. Software may accordingly configure a processor, for example, to constitute a particular hardware-implemented module at one instance of time and to constitute a different hardware-implemented module at a different instance of time.

[00124] Hardware-implemented modules may provide information to, and/or receive information from, other hardware-implemented modules. Accordingly, the described hardware-implemented modules may be regarded as being communicatively coupled. Where multiple of such hardware-implemented modules exist contemporaneously, communications may be achieved through signal transmission (e.g., over appropriate circuits and buses) that connect the hardware-implemented modules. In embodiments in which multiple hardware-implemented modules are configured or instantiated at different times, communications between such hardware-implemented modules may be achieved, for example, through the storage and retrieval of information in memory structures to which the multiple hardware-implemented modules have access. For example, one hardware-implemented module may perform an operation, and may store the output of that operation in a memory device to which it is communicatively coupled. A further hardware-implemented module may then, at a later time, access the memory device to retrieve and process the stored output. Hardware-implemented modules may also initiate communications with input or output devices.

[00125] While the present disclosure has been described with reference to various embodiments, it should be understood that these embodiments are illustrative and that the scope of the disclosure is not limited to them. Many variations, modifications, additions, and improvements are possible. More generally, embodiments in accordance with the present disclosure have been described in the context of particular implementations. Functionality may be separated or combined in blocks differently in various embodiments of the disclosure or described with different terminology. These and other variations, modifications, additions, and improvements may fall within the scope of the disclosure as defined in the claims that follow.

[00126] It should be understood from the foregoing that, while particular embodiments have been illustrated and described, various modifications can be made thereto without departing from the spirit and scope of the invention as will be

IPR2025-01398
CrowdStrike EX1002 Page 126

apparent to those skilled in the art. Such changes and modifications are within the scope and teachings of this invention as defined in the claims appended hereto.

IPR2025-01398
CrowdStrike EX1002 Page 127

CLAIMS

What is claimed is:

1. A method for assessing a likelihood of exploitation of software vulnerabilities, comprising:

utilizing a processor in operable communication with at least one memory for storing instructions that are executed the processor to perform operations of:

accessing a plurality of datasets associated with a predetermined set of data sources;

accessing training data comprising features and class labels associated with the features from the plurality of datasets;

applying learning algorithms to the training data to generate classification models that are configured to predict class labels defining a likelihood of exploitation of respective software vulnerabilities;

accessing one or more features associated with a software vulnerability; and

computing, by applying the one or more features to the classification model, a class label defining one or more values defining a likelihood of exploitation associated with the software vulnerability.

2. The method of claim 1, further comprising generating a plurality of estimation outputs based on the one or more values to derive an overall quantitative score.

3. The method of claim 1, wherein the plurality of datasets include vulnerability data for vulnerabilities that are publically disclosed and obtaining exploits data for exploits that were used in real world attacks.

4. The method of claim 3, further comprising:

aligning the exploits data with the vulnerability data; and

IPR2025-01398
CrowdStrike EX1002 Page 128

cleaning the exploits data of noise and predetermined portions of the exploits data that is irrelevant to associated software vulnerabilities.

5. The method of claim 1, wherein certain features correspond to a known vulnerability obtained from the plurality of datasets.

6. The method of claim 1, further comprising testing the classification models by applying additional training data and one or more algorithms and evaluation metrics to optimize the classification models until the classification models compute the likelihood of exploitation according to a predefined error rate.

7. The method of claim 1, further comprising vectorizing text features derived from the plurality of datasets using term frequency-inverse document frequency to create a vocabulary of associated words.

8. The method of claim 1, further comprising:

sorting vulnerabilities associated with the plurality of datasets according to time;

training the classification model using the training data, the training data defining a first subset of the plurality of datasets associated with a predetermined period of time; and

testing the classification model using a second subset of the plurality of datasets associated with the predetermined period of time.

9. The method of claim 1, further comprising computing mutual information from the plurality of datasets informative as to what information a given feature provides about another feature.

10. A computer-readable medium comprising instructions that cause a programmable processor to:

generate a learned function referencing features associated with a plurality of datasets defining software vulnerabilities and at least one machine learning algorithm; and

IPR2025-01398
CrowdStrike EX1002 Page 129

evaluate accuracy of the learned function by applying a portion of the plurality of datasets associated with software vulnerabilities to the learned function.

11. The computer-readable medium of claim 10 comprising additional instructions that cause the programmable processor to:

implement a random forest as part of the at least one machine learning algorithm that combines bagging for each tree with random feature selection at each node to split data utilized by the random forest, such that a result of implementing the random forest is an ensemble of decision trees each having their own independent opinion on class labels for a given disclosed vulnerability.

12. The computer-readable medium of claim 10 comprising additional instructions that cause the programmable processor to:

detect, from the plurality of datasets, vulnerabilities that appear before an associated exploitation date.

13. The computer-readable medium of claim 10 comprising additional instructions that cause the programmable processor to:

access features from the plurality of datasets that contain measures computed from social connections of users posting hacking-related content.

14. The computer-readable medium of claim 13 comprising additional instructions that cause the programmable processor to:

access features from the plurality of datasets that measure a centrality of the users in a social graph.

15. A computing device, configured via machine learning to apply a learned function derived from at least one machine learning algorithm and a plurality of datasets associated with software vulnerabilities to data associated with a software vulnerability to estimate a likelihood of exploitation of the software vulnerability.

IPR2025-01398
CrowdStrike EX1002 Page 130

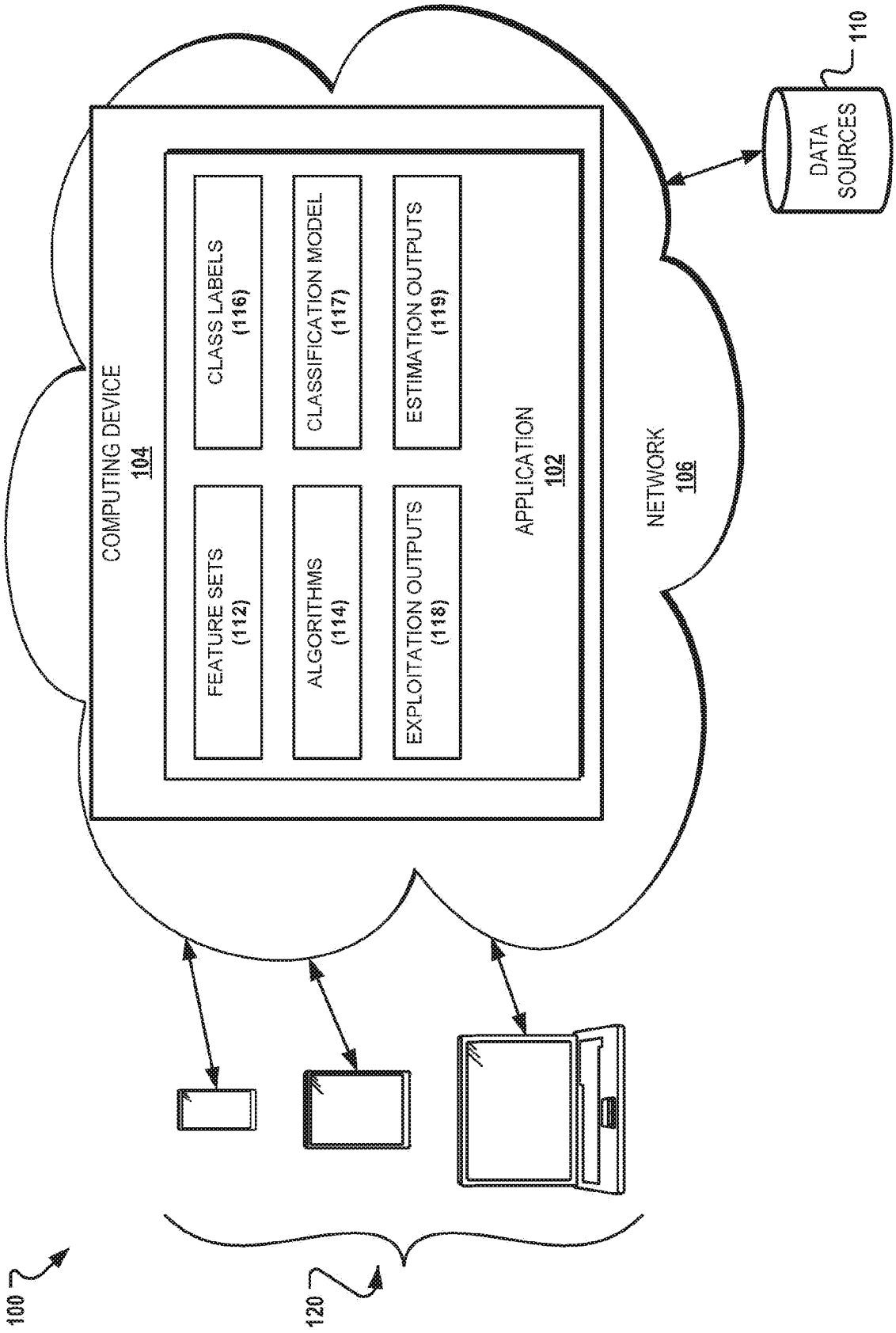

100
120
COMPUTING DEVICE
104
APPLICATION
102
FEATURE SETS
(112)
ALGORITHMS
(114)
EXPLOITATION OUTPUTS
(118)
CLASS LABELS
(116)
CLASSIFICATION MODEL
(117)
ESTIMATION OUTPUTS
(119)
NETWORK
106
DATA
SOURCES
110

FIG. 1A

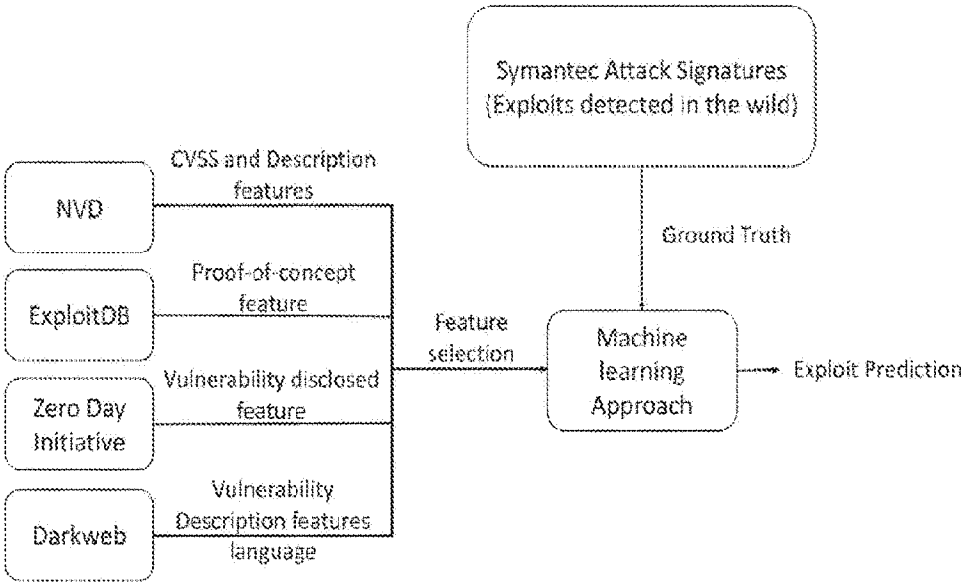


FIG. 1B



FIG. 2

IPR2025-01398
CrowdStrike EX1002 Page 132

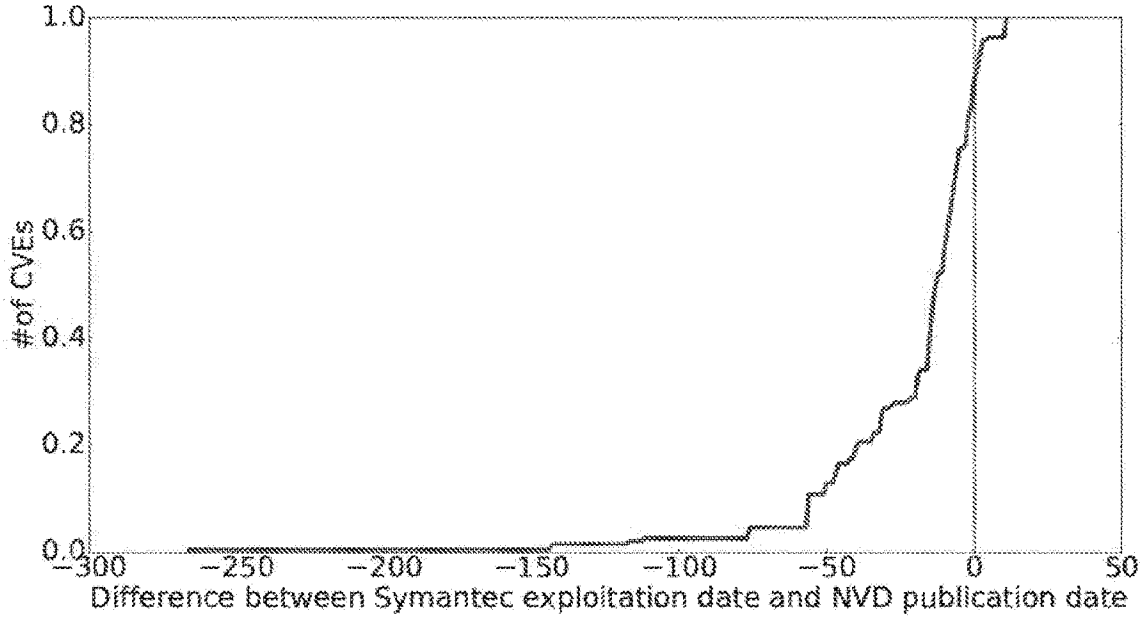

1.0
0.8
0.6
0.4
0.2
0.0
#of CVEs
−300
−250
−200
−150
−100
−50
0
50
Difference between Symantec exploitation date and NVD publication date

FIG. 3

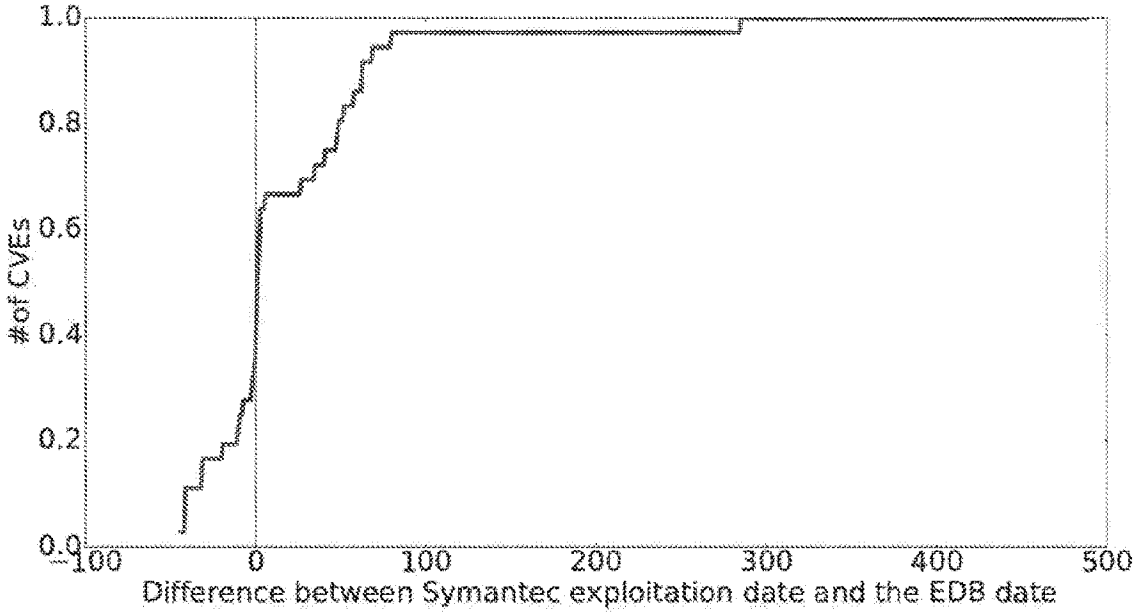

1.0
0.8
0.6
0.4
0.2
0.0
#of CVEs
−100
0
100
200
300
400
500
Difference between Symantec exploitation date and the EDB date

FIG. 4

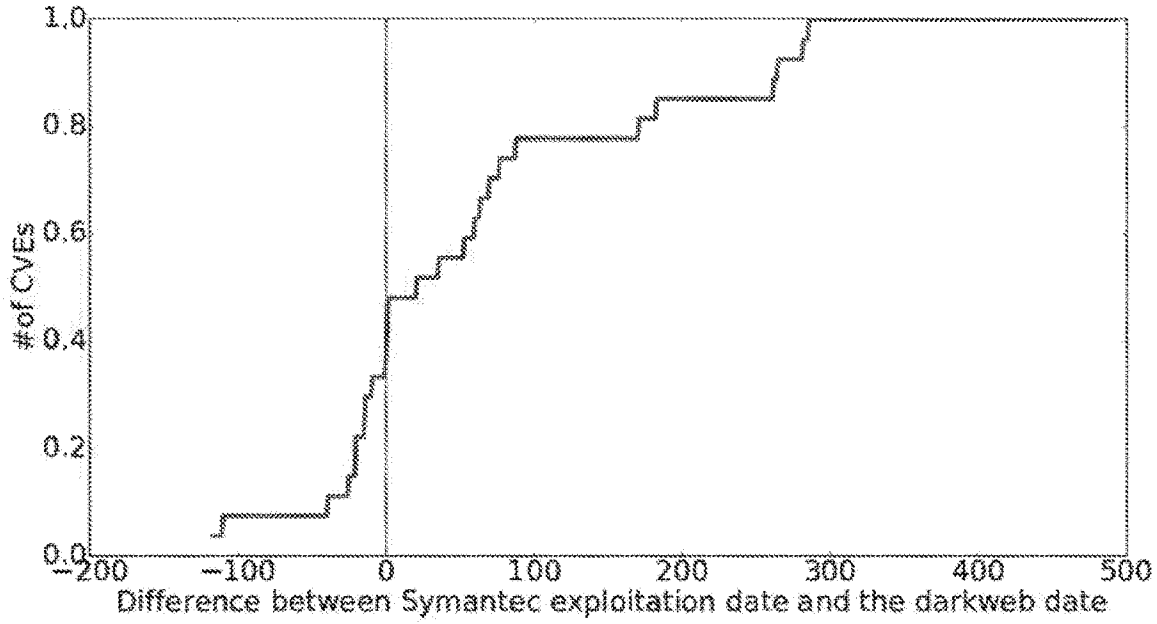

1.0
0.8
0.6
0.4
0.2
0.0
#of CVEs
−200
−100
0
100
200
300
400
500
Difference between Symantec exploitation date and the darkweb date

FIG. 5

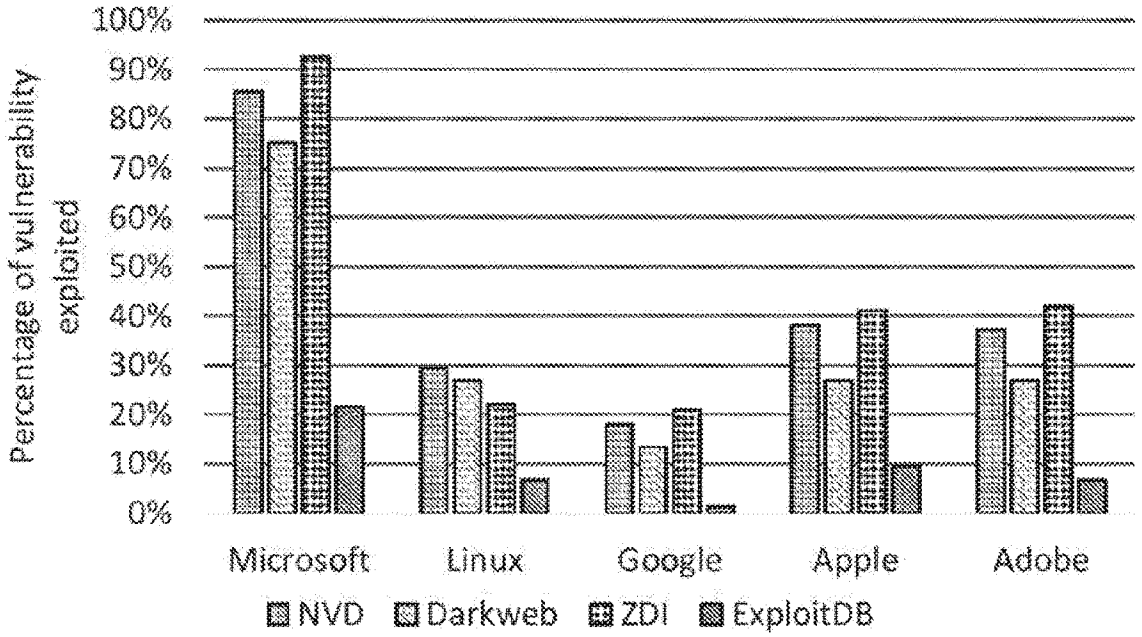

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percentage of vulnerability exploited
Microsoft
Linux
Google
Apple
Adobe
NVD
Darkweb
ZDI
ExploitDB

FIG. 6

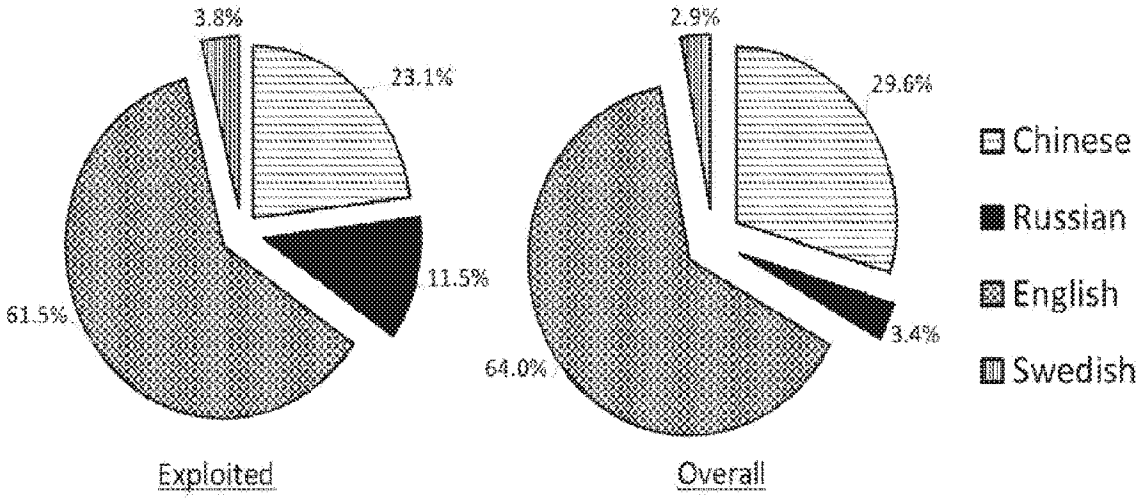


FIG. 7

IPR2025-01398
CrowdStrike EX1002 Page 135

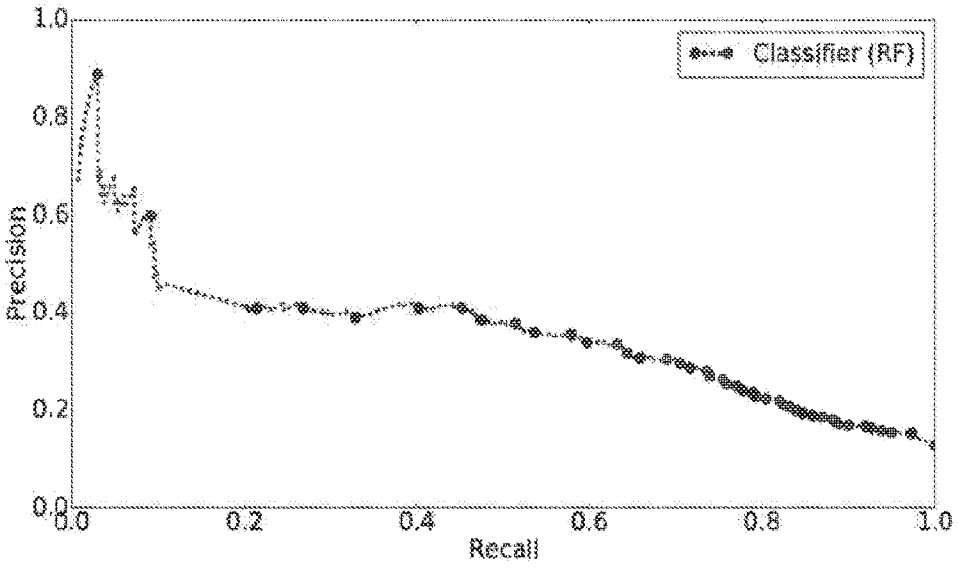


FIG. 8

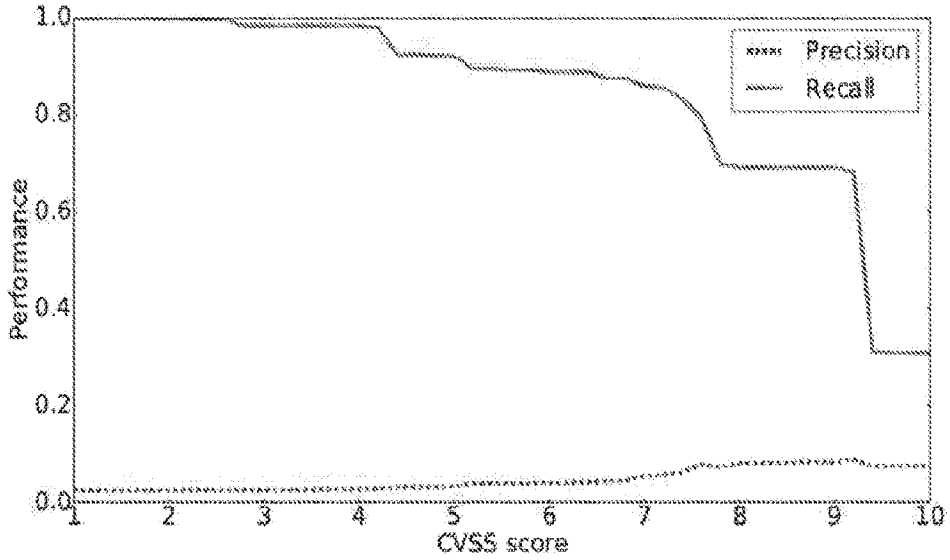


FIG. 9

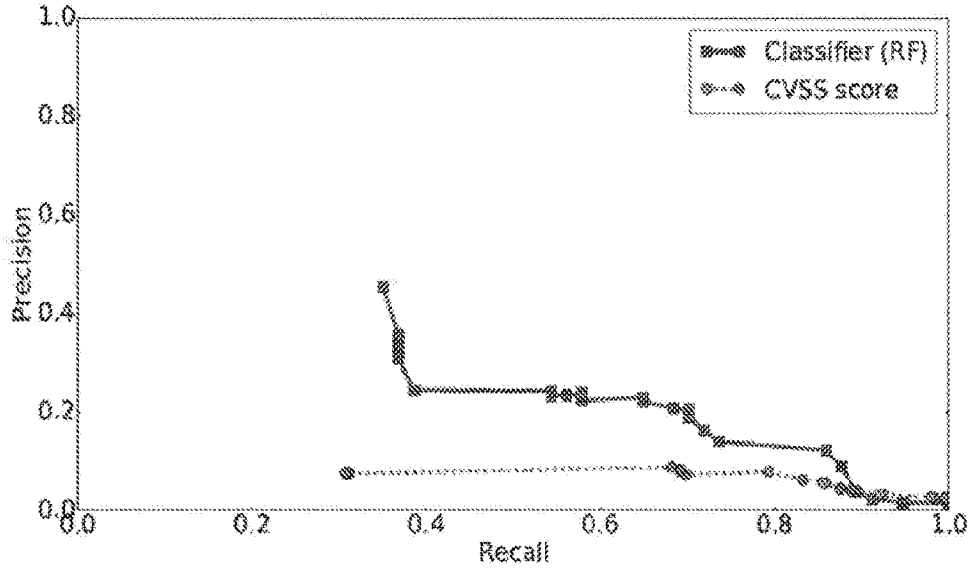


FIG. 10

IPR2025-01398
CrowdStrike EX1002 Page 136

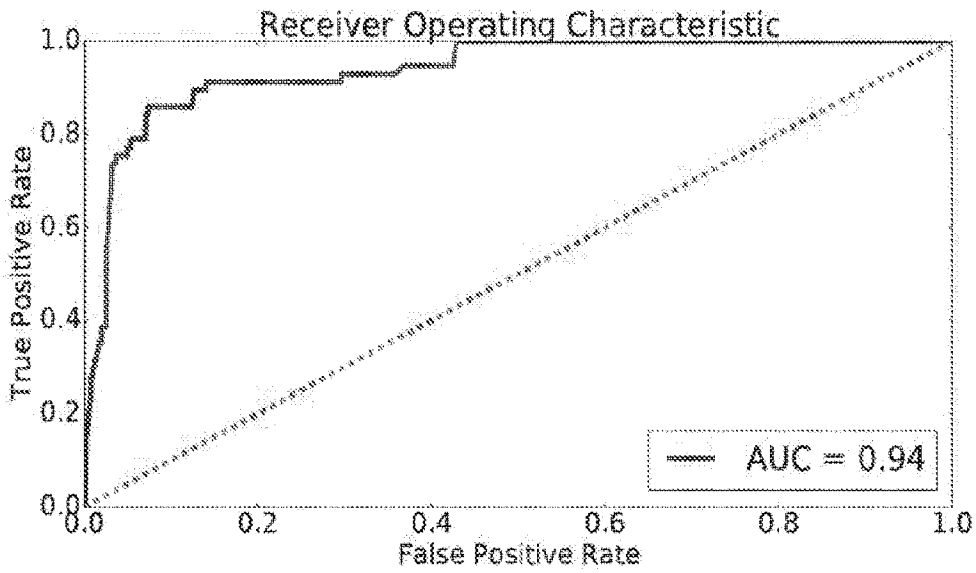


FIG. 11

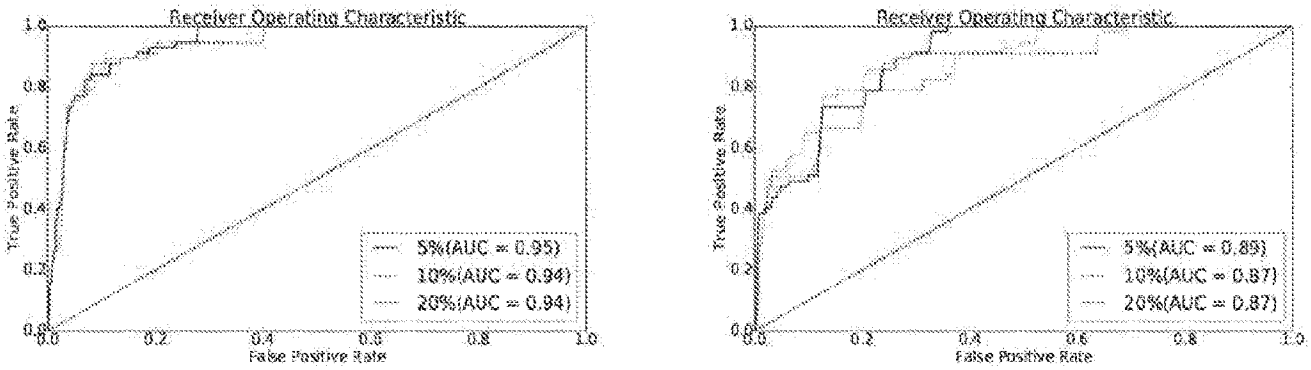


FIG. 12A

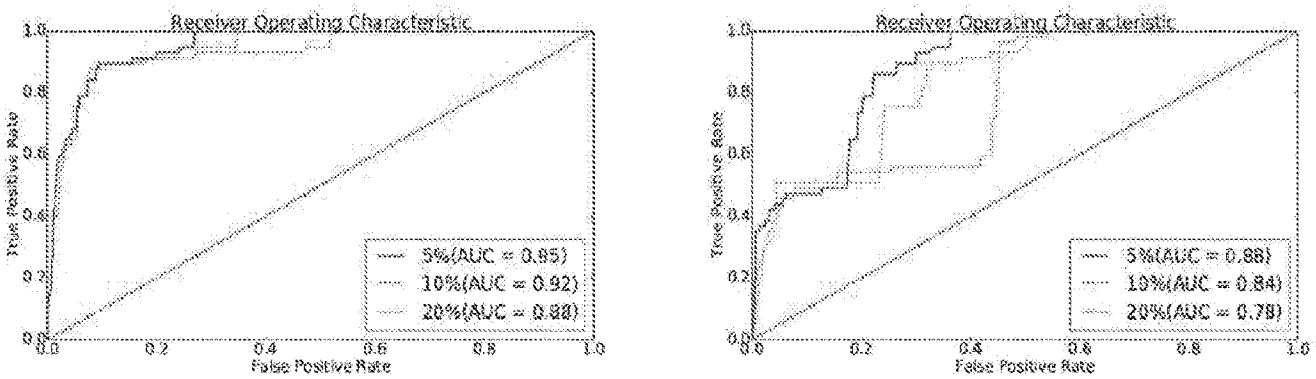


FIG. 12B

IPR2025-01398
CrowdStrike EX1002 Page 137

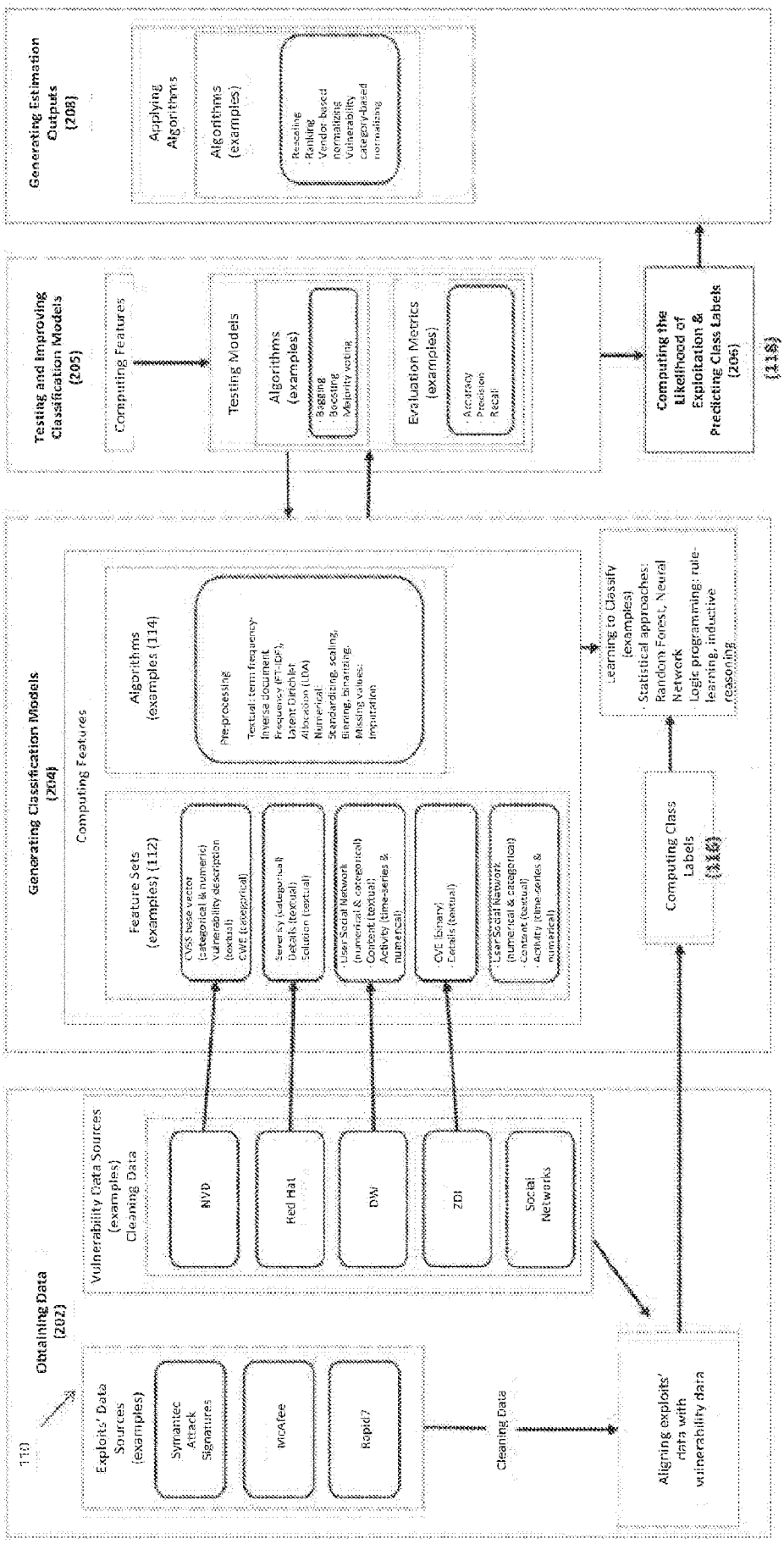


FIG. 13

IPR2025-01398
CrowdStrike EX1002 Page 138

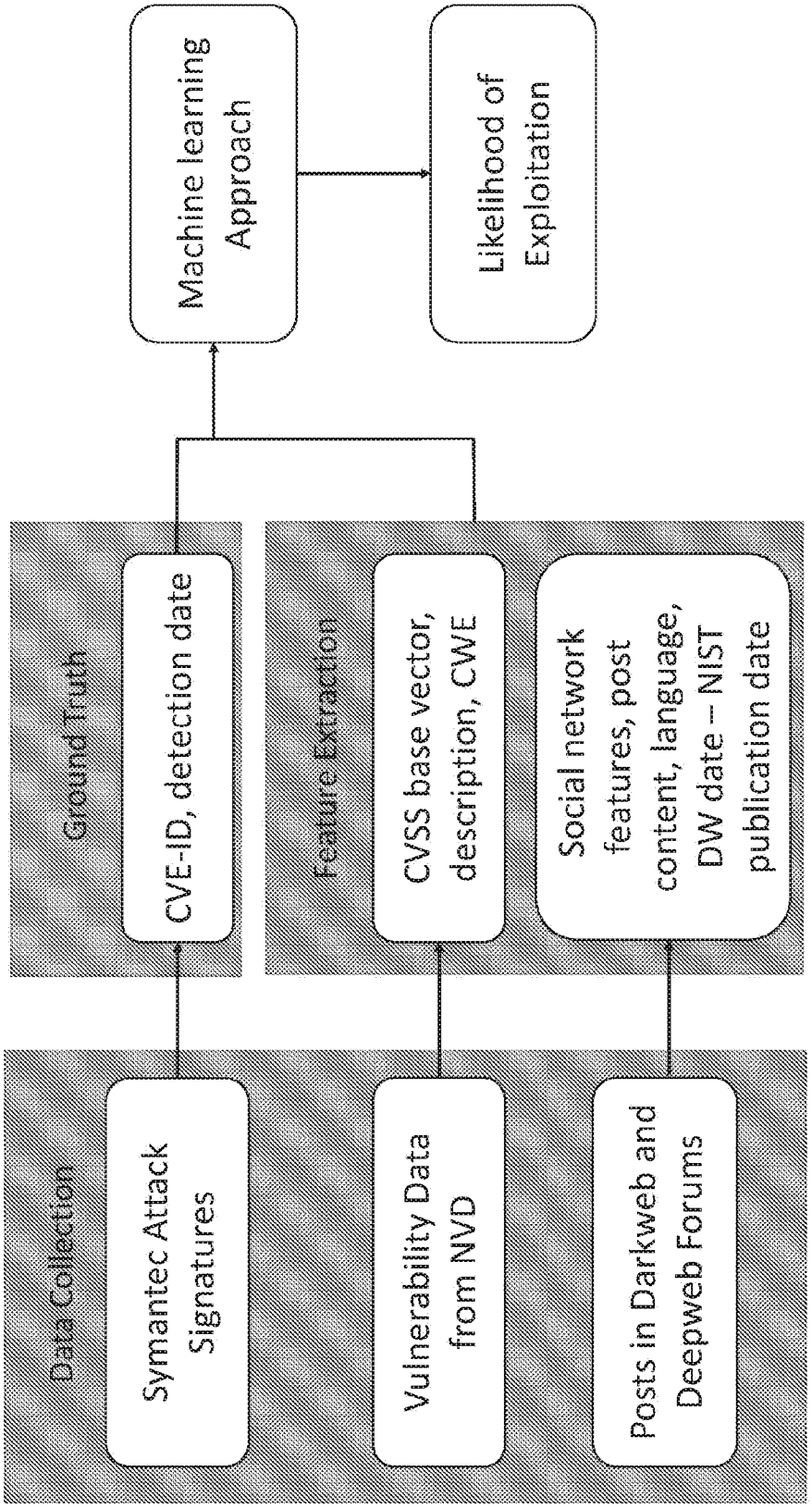

Machine learning Approach
Likelihood of Exploitation
Ground Truth
CVE-ID, detection date
Feature Extraction
CVSS base vector, description, CWE
Social network features, post content, language, DW date – NIST publication date
Data Collection
Symantec Attack Signatures
Vulnerability Data from NVD
Posts in Darkweb and Deepweb Forums

FIG. 14

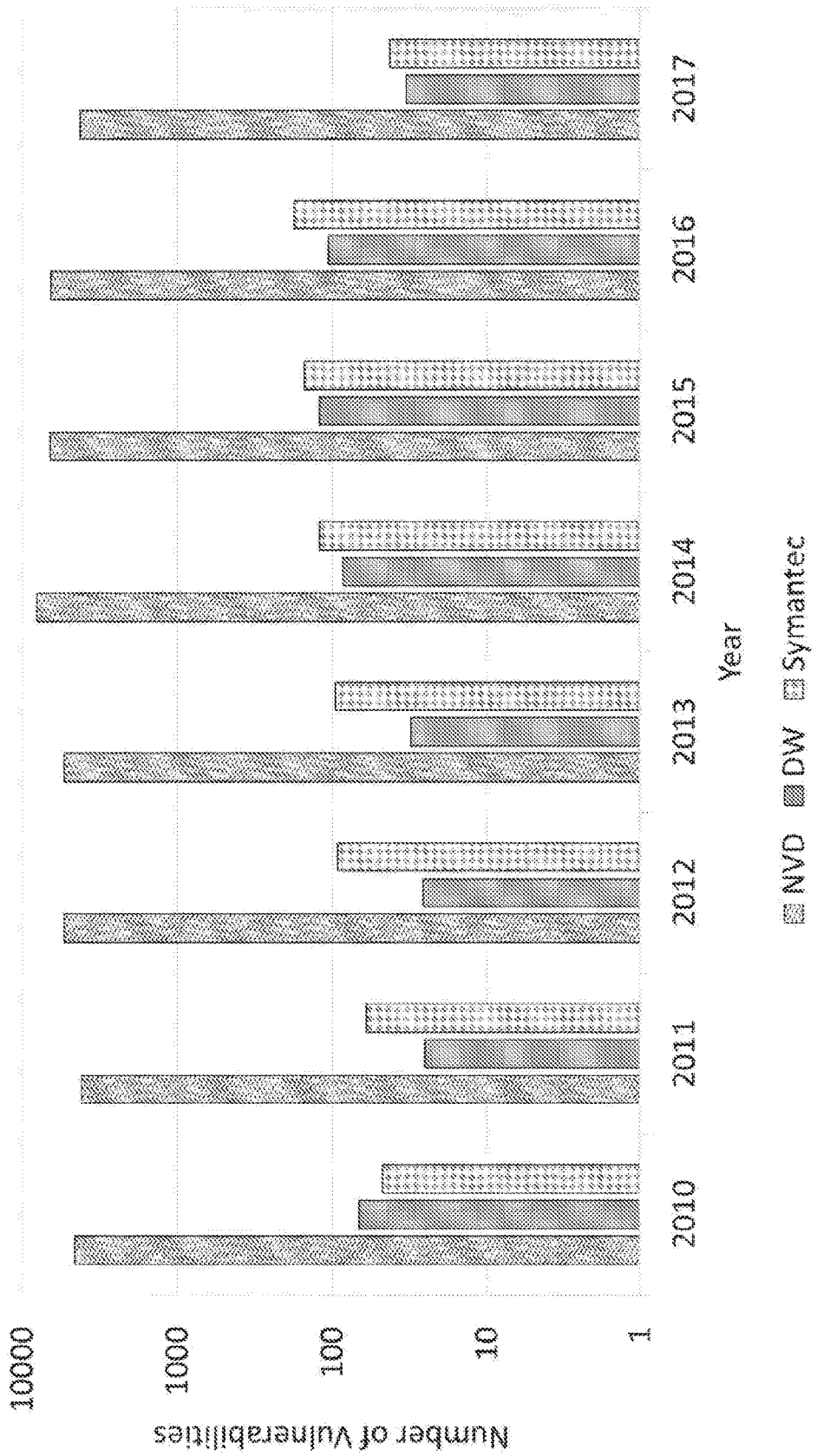

Number of Vulnerabilities
10000
1000
100
10
1
2010
2011
2012
2013
2014
2015
2016
2017
Year
NVD
DW
Symantec

FIG. 15

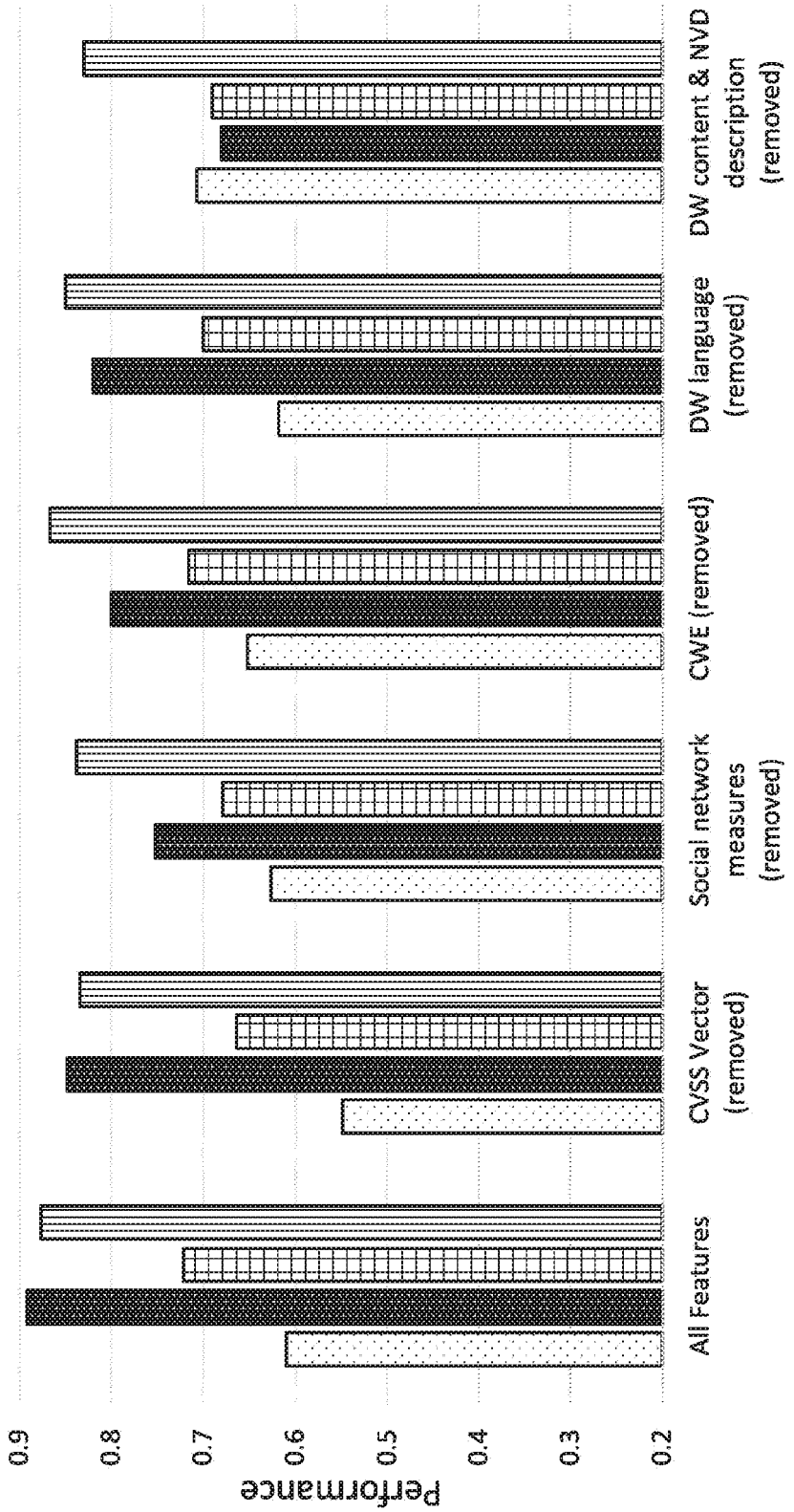

Performance
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
All Features
CVSS Vector (removed)
Social network measures (removed)
CWE (removed)
DW language (removed)
DW content & NVD description (removed)
Precision
Recall
F1
AUC

FIG. 16

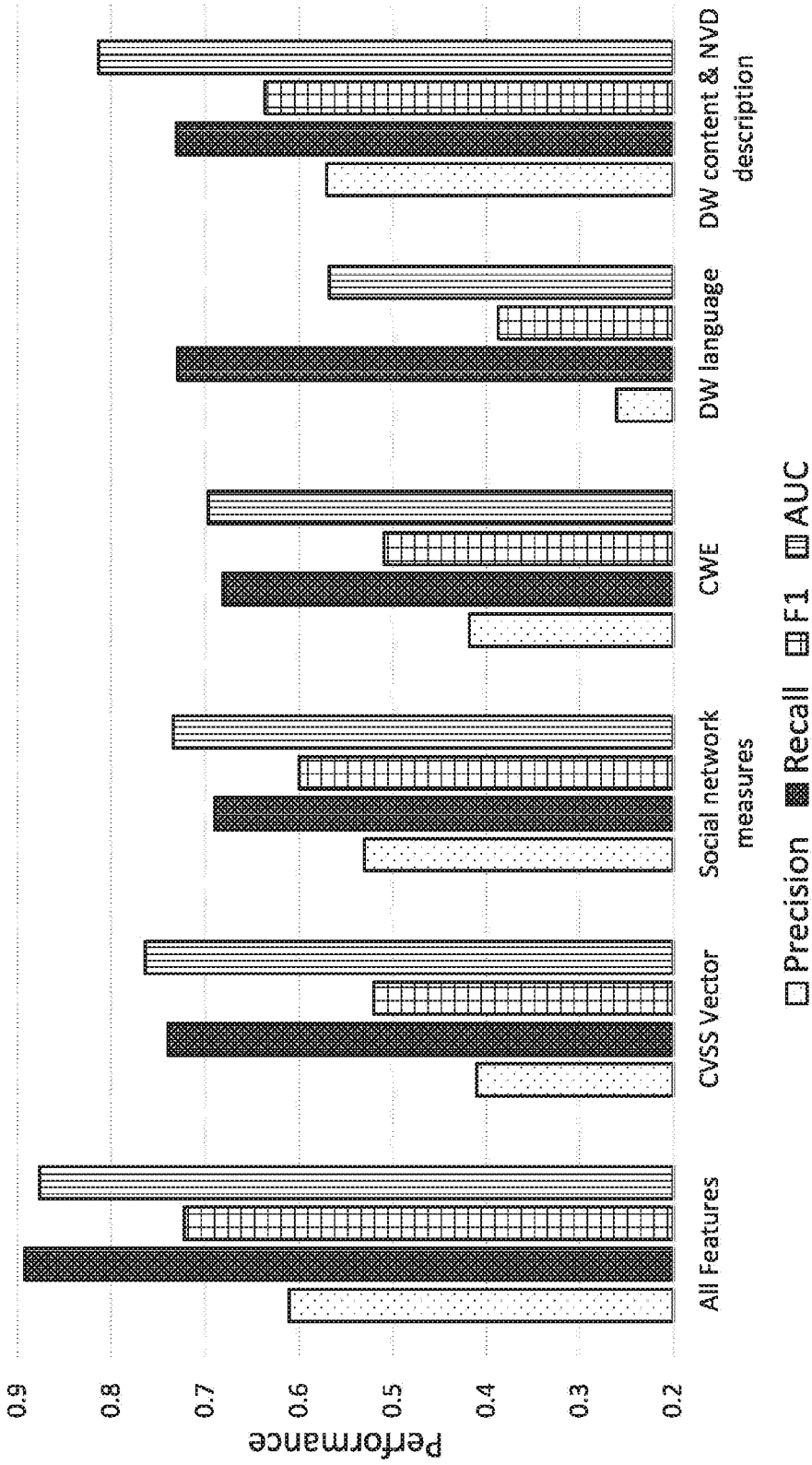

Performance
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
All Features
CVSS Vector
Social network measures
CWE
DW language
DW content & NVD description
Precision
Recall
F1
AUC

FIG. 17

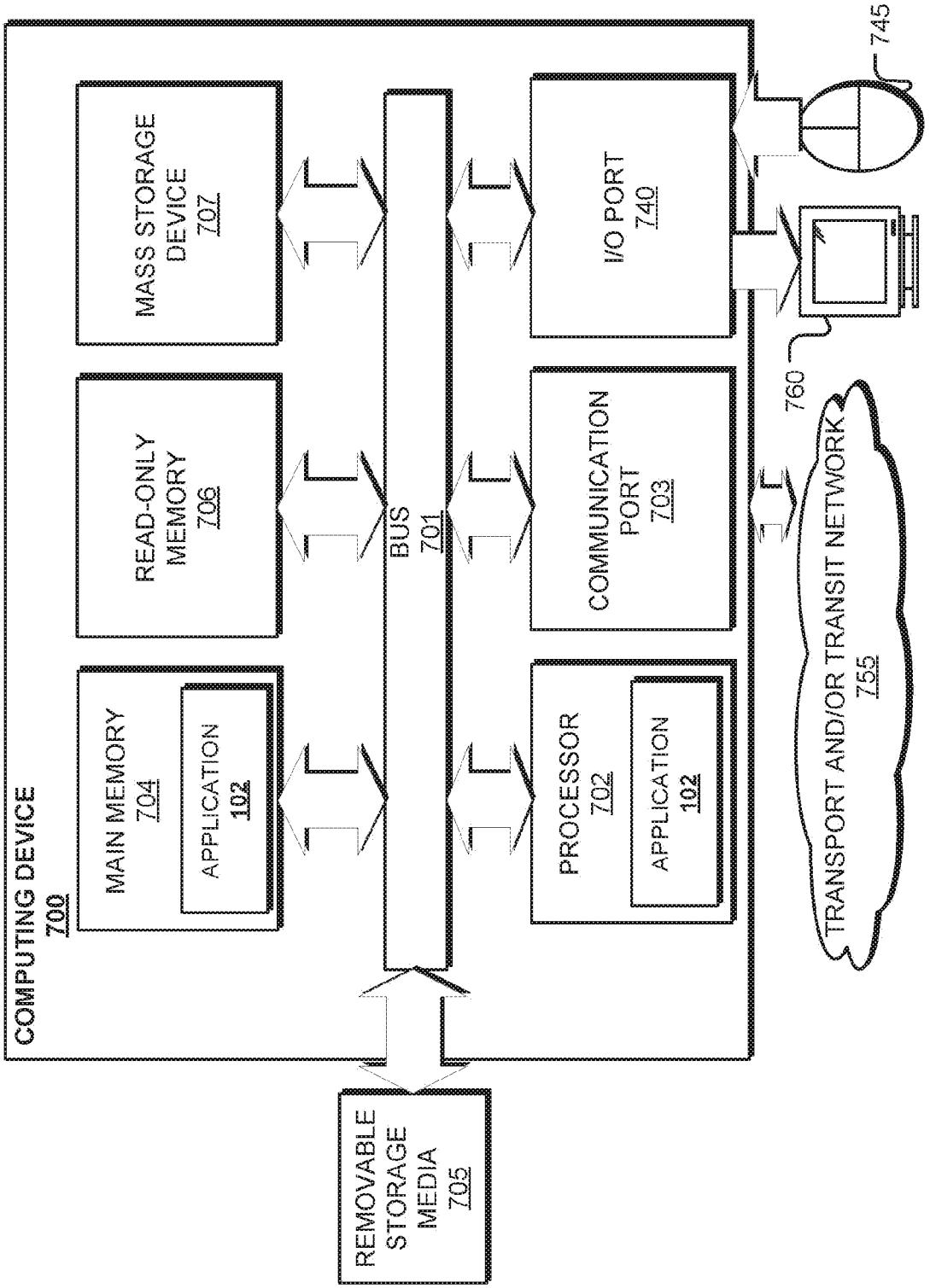

COMPUTING DEVICE
700
MAIN MEMORY
704
APPLICATION
102
READ-ONLY
MEMORY
706
MASS STORAGE
DEVICE
707
BUS
701
PROCESSOR
702
APPLICATION
102
COMMUNICATION
PORT
703
I/O PORT
740
745
760
TRANSPORT AND/OR TRANSIT NETWORK
755
REMOVABLE
STORAGE
MEDIA
705

FIG. 18

International application No.
PCT/US 18/57812

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 11/00, G06F 12/16, G06F 12/14 (2019.01)
CPC - G06F 21/577, H04L 63/1433, H04L 63/1416, H04L 63/20, H04L 63/1408

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History Document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
See Search History Document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
See Search History Document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X -- Y	US 2017/0214710 A1 (PALANTIR TECHNOLOGIES INC.) 27 July 2017 (27.07.2017), entire document, especially abstract and para [0074]-[0075], [0135], [0165], [0208]-[0216], [0227], Figs. 3-8.	1-6, 8-9, 15 ---------- 7
Y	US 2013/0097706 A1 (Titonis et al.) 18 April 2013 (18.04.2013), entire document, especially abstract and para [0151], [0263], claim 19.	7

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 January 2019 (28.01.2019)	21 FEB 2019

Name and mailing address of the ISA/US	Authorized officer:
Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

International application No.
PCT/US 18/57812

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

(See extra sheet)

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-9, 15

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

IPR2025-01398
CrowdStrike EX1002 Page 145

International application No.
PCT/US 18/57812

In continuation of Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet):

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1.

Group I: Claims 1-9, 15, directed to assessing a likelihood of exploitation of software vulnerabilities.

Group II: Claims 10-14, directed to generating a learned function referencing features associated with software vulnerabilities, and evaluating accuracy of the learned function.

The groups of inventions listed above do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

The special technical feature of the Group I invention is applying machine learning algorithms/functions and a plurality of datasets, to estimate/predict a likelihood of exploitation of the software vulnerability, not required by Group II.

The special technical feature of the Group II invention is evaluating accuracy of the learned function by applying a portion of the plurality of datasets associated with software vulnerabilities to the learned function, not required by Group I.

Groups I-II share the technical features of, in various combination, utilizing/applying a learned function/machine learning, in conjunction with a machine learning algorithm, a plurality of datasets, in conjunction with data/features associated with software vulnerabilities.

However, these shared technical features fail to represent a contribution over the prior art of US 2017/0214710 A1 to Palantir Technologies Inc. (hereinafter 'Palantir'), which discloses utilizing/applying a learned function/machine learning, in conjunction with a machine learning algorithm, a plurality of datasets, in conjunction with data/features associated with software vulnerabilities (para [0135], [0165], [0208]-[0216], [0227] - "Compromise likelihood may consider linkages (e.g. proximity to insecure parts of the network like the demilitarized zone of the network), attributes (e.g. software version) for a given node, and/or an academic theory like attack graphs in computing a compromise likelihood for a node. This information can be provided to the machine learning model, described above, and the system can obtain a probability identifying a chance the node will get compromised"; "the system can provide the determined compromise vulnerability and compromise value to a machine learning algorithm trained on labeled data to determine a network device risk value for the network device").

Thus, the inventions listed as Groups I-II lack unity of invention because they do not share a same or corresponding special technical feature providing a contribution over the prior art.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 August 2019 (15.08.2019)

WIPO | PCT

(10) International Publication Number
WO 2019/157335 A1

(51) International Patent Classification:
G06F 15/173 (2006.01) *G06F 21/50* (2013.01)
G06F 16/24 (2019.01)

(21) International Application Number:
PCT/US2019/017287

(22) International Filing Date:
08 February 2019 (08.02.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/628,196 08 February 2018 (08.02.2018) US

(71) Applicant *(for all designated States except US)*: **ARIZONA BOARD OF REGENTS ON BEHALF OF ARIZONA STATE UNIVERSITY** [US/US]; 699 S. Mills Avenue, Suite 601, Room 691AA, Tempe, Arizona 85281 (US).

(72) Inventors; and

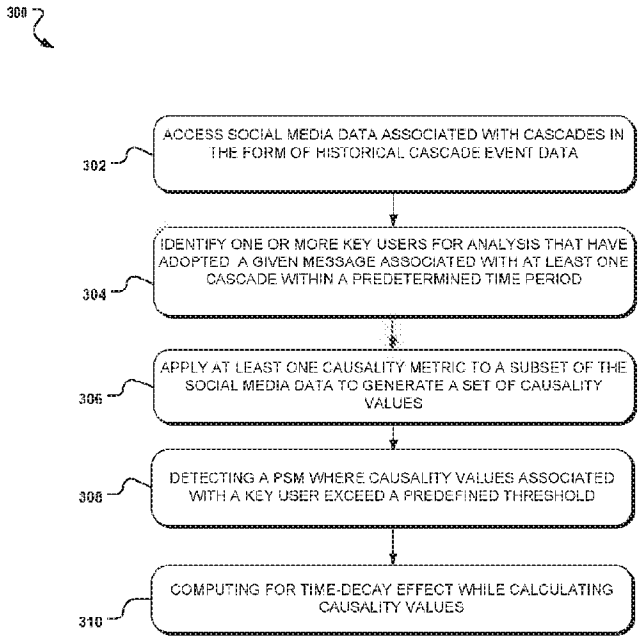
(71) Applicants *(for US only)*: **SHAABANI, Elham** [IR/US]; c/o ARIZONA BOARD OF REGENTS ON BEHALF OF ARIZONA STATE UNIVERSITY, 699 S. Mills Avenue, Suite 601, Room 691AA, Tempe, Arizona 85281 (US). **SHAKARIAN, Paulo** [US/US]; c/o ARIZONA BOARD OF REGENTS ON BEHALF OF ARIZONA STATE UNIVERSITY, 699 S. Mills Avenue, Suite 601, Room 691AA, Tempe, Arizona 85281 (US). **ALVARI, Hamidreza** [IR/US]; c/o ARIZONA BOARD OF REGENTS ON BEHALF OF ARIZONA STATE UNIVERSITY, 699 S. Mills Avenue, Suite 601, Room 691AA, Tempe, Arizona 85281 (US). **GUO, Ruocheng** [CN/US]; c/o ARIZONA BOARD OF REGENTS ON BEHALF OF ARIZONA STATE UNIVERSITY, 699 S. Mills Avenue, Suite 601, Room 691AA, Tempe, Arizona 85281 (US).

(74) Agent: **FONTES, Timothy** et al.; POLSINELLI PC, 105 West Vandalia Street, Suite 400, Edwardsville, Illinois 62025 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM,

(54) Title: SYSTEMS AND METHODS FOR DETECTING PATHOGENIC SOCIAL MEDIA ACCOUNTS WITHOUT CONTENT OR NETWORK STRUCTURE

FIG. 20



(57) Abstract: Embodiments of a system and methods for detecting social media designed to spread malicious information to "viral" proportions are disclosed. Historical cascade event data from preselected social media accounts as well as information from related accounts is applied to one or more causality metrics to generate a set of causality values. Causality values are further refined and analyzed to determine how casual a user is with respect to a cascade as opposed to other similar users.

WO 2019/157335 A1

[Continued on next page]

IPR2025-01398
CrowdStrike EX1002 Page 147

AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

SYSTEMS AND METHODS FOR DETECTING PATHOGENIC SOCIAL MEDIA ACCOUNTS WITHOUT CONTENT OR NETWORK STRUCTURE

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This is a PCT application that claims benefit to U.S. provisional application serial number 62/628,196 filed on February 8, 2018 which is incorporated by reference in its entirety.

FIELD

[0002] The present disclosure generally relates to identifying pathogenic social media accounts, and in particular to systems and methods for detecting pathogenic social media accounts without leveraging network structure, cascade path information, content, or user's information.

BACKGROUND

[0001] The spread of harmful misinformation in social media is a pressing problem. Social media accounts that have the capability of spreading such information to viral proportions are referred to as "Pathogenic Social Media" (PSM) accounts. These accounts can include terrorist supporter accounts, water armies, and fake news writers. In addition, these organized groups/accounts spread messages regarding certain topics. Such accounts might be operated by multiple people that tweet/retweet through multiple accounts to promote/degrade an idea, which can influence public opinion. As such, identifying PSM accounts has important applications to countering extremism, the detection of water armies, and fake news campaigns. In Twitter, many of these accounts have been found to be social bots. The PSM accounts that propagate information are key to a malicious information campaign and detecting such accounts is critical to understanding and stopping such campaigns. However, this is difficult in practice. Existing methods rely on social media message content (e.g. text, image, video, audio, hashtag, URL, or user mentions, etc.), social media user profile information (e.g. user name, photo, URL, number of tweets/retweets, location, etc.), social or network structure (e.g. number of followers, number of followees, actual list of followers, actual list of followees, number of friends, actual friends list, betweenness centrality, pagerank centrality,

IPR2025-01398
CrowdStrike EX1002 Page 149

closeness centrality, and in-degree and out-degree centralities, star and clique networks associated with users, average clustering coefficients of each user, etc.), or combinations thereof. However, reliance on information of this type leads to two challenges. First, collection, storage, and processing of message content, user profile, and/or network structure requires significantly more computing CPU, memory, and network bandwidth. Second, the use of content often necessitates the training of a new model for previously unobserved topics. For example, PSM accounts taking part in elections in the U.S. and Europe will likely leverage different types of content. Third, network structure information is not always available. For example, the FACEBOOK® API does not make this information available without permission of the users (which is likely a non-starter for identifying PSM accounts).

[0002] Early detection of PSM accounts is critical as these accounts are likely to be key users related to the formation of harmful malicious information campaigns. This is a challenging task for social media authorities for three reasons. First, it is not always easy to manually shut down these accounts. Despite efforts to suspend these accounts, many of them simply return to social media with different accounts. Second, for the most part, the available data is imbalanced and social network structure, which is at the core of many techniques are not readily available. Third, PSM accounts often seek to utilize and cultivate large number of online communities of passive supporters to spread as much harmful information as they can. Consequently, extra efforts need to be dedicated to proposing capabilities that could be deployed by social network firms to counter PSM accounts, regardless of the underlying social network platform.

[0001] It is with these observations in mind, among others, that various aspects of the present disclosure were conceived and developed.

BRIEF DESCRIPTION OF THE DRAWINGS

[0002] FIGS. **1A-1D** show a hypergraph of three cascades, according to aspects of the present disclosure;

[0003] FIGS. **2A-2B** are graphical representations of the number of cascades vs. cascade size, according to aspects of the present disclosure;

[0004] FIGS. **3A-3C** are graphical representations showing the cumulative distribution of the number of times a user has participated in cascades, according to aspects of the present disclosure;

IPR2025-01398
CrowdStrike EX1002 Page 150

[0005] FIGS. **4A-4D** are graphical representations showing the distribution between active and inactive users, according to aspects of the present disclosure;

[0006] FIG. **5** shows a log-log distribution of cascades vs. cascade size (top) and cumulative distribution of duration of cascades (bottom);

[0007] FIG. **6** shows the total number of inactive users in each cascade;

[0008] FIG. **7** shows the number of paired URLs posted by the PSM and normal users in the dataset. Note that the normal users in the dataset are higher than the PSM accounts.

[0009] FIG. **8** illustrates the Hawkes Process. Events induce impulse on other processes and cause child events. Background event in e_0 induces impulse on responses on processes e_1 and e_2.

[0010] FIG. **9** is a graphical representation showing a comparison between threshold-based and label propagation selection approaches, according to aspects of the present disclosure;

[0011] FIG. **10** shows the estimated weights for all paired URLs for (top) PSM accounts, and (bottom) normal users. Among all social media platforms and mainstream and alternative news sources, URLs shared from facebook.com and alternative news media had largest impact on dissemination of malicious messages.

[0012] FIG. **11** is an illustration of a proposed metric, according to aspects of the present disclosure;

[0013] FIGS. **12A-12D** are illustrations of box plots of the distributions of the decay-based metrics for active and inactive users, according to aspects of the present disclosure;

[0014] FIGS. **13A-13D** are graphical representations showing plots of distributions of users using decay-based metrics according to aspects of the present disclosure;

[0015] FIGS. **14A-14B** are graphical representations showing the precision and recall for each classifier, according to aspects of the present disclosure;

[0016] FIGS. **15A-15B** are graphical representations showing the F1 score and AUC results for each classifier, according to aspects of the present disclosure.

IPR2025-01398
CrowdStrike EX1002 Page 151

[0017] . FIGS. **16A-16B** show cumulative true positives and false positives based on the available training data.

[0018] FIG. **17** compares diskspace requirements from existing approaches due to content storage as opposed to specifications according to embodiments of the present novel concept, which primarily uses action logs.

[0019] FIG. **18** is a network diagram for illustrating generation of the social media data described herein, according to aspects of the present disclosure.

[0020] FIG. **19** is a system diagram illustrating one possible non-limiting embodiment for transforming the social media data generated in FIG. **18** to causality values for detecting PSM accounts using the novel methods and functionality described herein, according to aspects of the present disclosure.

[0021] FIG. **20** is a process flow illustrating possible steps of implementing the functionality described herein and shown in FIG. **19**, according to aspects of the present disclosure.

[0022] FIG. **21** is a schematic diagram of a non-limiting computing system for implementing the functionality described herein, according to aspects of the present disclosure.

[0023] Corresponding reference characters indicate corresponding elements among the view of the drawings. The headings used in the figures do not limit the scope of the claims.

DETAILED DESCRIPTION

GENERAL DESCRIPTION OF INVENTION

[0024] As described herein, addressing the viral spread of misinformation poses various technical problems. For example, some technologies may rely upon social media message content, social media user profile information, social structure, or combinations thereof and therefore require excess computing power, memory, and network bandwidth. The present novel concept provides a technical solution to these technical problems by defining a sophisticated computational framework that determines how causal a user is with respect to a cascade by comparison to others using historical cascade event data. The concept further differs from previous approaches by focusing on malicious social media accounts that make messages "go viral". The framework may supplement but does not require the analysis of social media message content, social media user profile

IPR2025-01398
CrowdStrike EX1002 Page 152

information, or social to identify PSM accounts. Rather, the concept relies primarily on the evaluation of how likely a given social media account is to cause a message to spread virally. Thus, the present disclosure involves significantly less CPU, memory, and network bandwidth, and at the same time, provides a more accurate and efficient system for detecting malicious accounts. Further, because the concept does not rely upon social media content, it can potentially detect pathogenic social media accounts used to spread new types of malicious information where the content has differed greatly from training data. This approach is potentially complementary to content based methods.

TECHNICAL APPROACH

Technical Preliminaries

[0025] Throughout this disclosure cascades shall be represented as an "action log" (*Actions*) of tuples where each tuple $(u, m, t) \in Actions$ corresponds with a user $u \in U$ posting message $m \in M$ at time $t \in T$. It is assumed that set M includes posts/repost of a certain original tweet or message. For a given message, the first occurrence of each user is only considered. The present disclosure defines $Actions_m$ as a subset of *Actions* for a specific message m. Formally, it is defined as $Actions_m = \{(u', m', t') \in Actions \text{ s.t. } m' = m\}$.

Definition 1. (*m*-participant). *For a given $m \in M$, user i is an m-**participant** if there exists t such that $(i, m, t) \in Actions$.*

[0026] Note that the users posting tweet/retweet in the early stage of cascades are the most important ones since they play a significant role in advertising the message and making it viral. For a given $m \in M$, an m-participant i "precedes" m-participant j if there exists $t < t'$ where $(i, m, t), (j, m, t') \in Actions$. Thus, key *users* are defined as a set of users adopting a message in the early stage of its life span. These users appear in the "beginning" of a cascade, before specified users join the cascade. The present disclosure defines *key user* as follows:

Definition 2. (Key User). *For a given message m, m-participant i, and $Actions_m$, wherein user i is a **key user** iff user i precedes at least ϕ fraction of m-participants*

IPR2025-01398
CrowdStrike EX1002 Page 153

(formally: $|Actions_m| \times \phi \leq |\{j | \exists t' : (j, m, t') \in Actions_m \wedge t' > t\}|$, $(i, m, t) \in Actions_m$*), where* $\phi \in (0, 1)$.

[0027] The notation $|\cdot|$ denotes the cardinality of a set. All messages are not equally important. That is, only a small portion of them become popular. These *viral messages* are defined as follows:

Definition 3. (Viral Messages). *For a given threshold* θ*, a message* $m \in M$ *is* ***viral*** *iff* $|Actions_m| \geq \theta$*.* M_{vir} *is used to denote the set of viral messages.*

[0028] The Definition 3 allows the computation of the prior probability of a message (cascade) going viral as follows:

$$\rho = \frac{|M_{vir}|}{|M|} \tag{1}$$

[0029] The probability of a cascade *m* going viral given some user *i* was involved is defined as:

$$p_{m|i} = \frac{|\{m \in M_{vir} \; s.t. \; i \; is \; a \; key \; user\}|}{|\{m \in M \; s.t. \; i \; is \; a \; key \; user\}|} \tag{2}$$

[0030] The present disclosure is also concerned with two other measures. First, the probability that two users *i* and *j* tweet or retweet viral post *m* chronologically, and both are key users. In other words, these two users are making post *m* viral.

$$p_{i,j} = \frac{|\{m \in M_{vir} | \exists t, t' \; where \; t < t' \; and \; (i, m, t), (j, m, t') \in Actions\}|}{|m \in M | \exists t, t' \; where \; (i, m, t), (j, m, t') \in Actions|} \tag{3}$$

[0031] Second, the probability that key user *j* tweets/retweets viral post *m* and user *i* does not tweet/retweet earlier than *j*. In other words, only user *j* is making post *m* viral.

$$p_{\neg i,j} = \frac{|\{m \in M_{vir} | \exists t' \; s.t. \; (j, m, t') \in Actions \; and \; \nexists t \; where \; t < t', \; (i, m, t) \in Actions\}|}{|\{m \in M | \exists t' \; s.t. \; (j, m, t') \in Actions \; and \; \nexists t \; where \; t < t', \; (i, m, t) \in Actions\}|} \tag{4}$$

[0032] Knowing the action log, the aim is to find a set of pathogenic social media (PSM) accounts. These users are associated with the early stages of

IPR2025-01398
CrowdStrike EX1002 Page 154

large information cascades and, once detected, are often deactivated by a social media firm. In the causal framework disclosed herein, a series of causality-based metrics is introduced for identifying PSM users.

Causal Framework

[0033] The present disclosure adopts a causal inference framework. In particular, the present disclosure expands upon the causal inference framework in two ways: (1) the casual framework addresses the problem of identifying PSM accounts; and (2) the present disclosure extends this single causal metric to a set of metrics. Multiple causality measurements provide a stronger determination of significant causality relationships. For a given viral cascade, the casual framework seeks to identify potential users who likely caused the cascade to go viral. An initial set of criteria is first required for such a causal user. This is done by instantiating the notion of Prima Facie causes to our particular use case below:

Definition 4. (Prima Facie Causal User). *A user u is a prima facie causal user of cascade m iff: User u is a key user of m, $m \in M_{vir}$, and $p_{m|u} > \rho$.*

[0034] For a given cascade *m*, the language *prima facie causal user* will often be used to describe user *i* as a prima facie cause for *m* to be viral. In determining if a given prima facie causal user is causal, other "related" users must be considered. In this disclosure, *i* and *j* are *m*-related if (1.) *i* and *j* are both prima facie causal users for *m*, (2.) *i* and *j* are both key users for *m*, and (3.) *i* precedes *j*. Hence, the set of "related users" for user *i* (denoted R(*i*)) are defined as follows:

$$R(i) = \{j \; s.t. \; j \neq i, \exists m \in M \; s.t. \; i, j \; are \; m-related\} \tag{5}$$

[0035] Therefore, $p_{i,j}$ in (3) is the probability that cascade *m* goes viral given both users *i* and *j*, and $p_{\neg i,j}$ in (4) is the probability that cascade *m* goes viral given key user *j* tweets/retweets it while key user *i* does not tweet/retweet *m* or precedes *j*. The idea is that if $p_{i,j} - p_{\neg i,j} > 0$, then user *i* is more likely a cause than *j* for *m* to become viral. The *Kleinberg-Mishra causality* ($\epsilon_{K\&M}$) is measured as the average of this quantity to determine how causal a given user *i* is as follows:

$$\epsilon_{K\&M}(i) = \frac{\sum_{j \in R(i)} (p_{i,j} - p_{\neg i,j})}{|R(i)|} \tag{6}$$

IPR2025-01398
CrowdStrike EX1002 Page 155

[0036] Intuitively, $\epsilon_{K\&M}$ measures the degree of causality exhibited by user i. In addition, it was found useful to include a few other measures. The *relative likelihood* causality (ϵ_{rel}) is introduced as follows:

$$\epsilon_{rel}(i) = \frac{\sum_{j \in R(i)} S(i,j)}{|R(i)|} \tag{7}$$

$$S(i,j) = \begin{cases} \left(\frac{p_{i,j}}{p_{\neg i,j}+\alpha}\right) - 1, & p_{i,j} > p_{\neg i,j} \\ 0, & p_{i,j} = p_{\neg i,j} \\ 1 - \left(\frac{p_{\neg i,j}}{p_{i,j}}\right), & \text{otherwise} \end{cases} \tag{8}$$

where α is infinitesimal. Relative likelihood causality metric assesses the relative difference between $p_{i,j}$ and $p_{\neg i,j}$. This assists in finding new users that may not be prioritized by $\epsilon_{K\&M}$. It was also found that if a user is mostly appearing after those with the high value of $\epsilon_{K\&M}$, then it is likely to be a PSM account. One can consider all possible combinations of events to capture this situation. However, this approach is computationally expensive. Therefore, Q(*j*) is defined as follows:

$$Q(j) = \{i : i, j \in R(i)\} \tag{9}$$

Consider the following example:

Example 1. *Consider two cascades (actions)* τ_1 = {A, B, C, D, E, F, G, H} *and* τ_2 = {N, M, C, A, H, V, S, T} *where the capital letters signify users. The aim is to relate key users while* ϕ = *0.5 (Definition 2). Table I shows the related users R(.) for each cascade. Note that the final set R(.) for each user, is the union of all sets from the cascades. Set Q(.) for the users of Table I are presented in Table II.*

TABLE I: Related users R(.) (5) of cascades τ_1 = {A, B, C, D, E, F, G, H} *and* τ_2 = {N, M, C, A, H, V, S, T}

User	$R\,\tau_1$	$R\,\tau_2$	R
A	{B, C, D, E, F}	{H, V}	{B, C, D, E, F, H, V}
B	{C, D, E, F}	{}	{C, D, E, F}
C	{D, E, F}	{A, H, V}	{A,D, E, F, H, V}
D	{E, F}	{}	{E, F}
E	{F}	{}	{F}
N	{}	{M, C, A, H, V}	{A, C, H, M, V}
M	{}	{C, A, H, V}	{A, C, H, V}

IPR2025-01398
CrowdStrike EX1002 Page 156

H	{}	{V}	{V}

TABLE II: Set Q(.) of users Table I in (9)

User	Total
A	{C, N, M}
B	{A}
C	{A, B, N, M}
D	{A, B, C}
E	{A, B, C, D}
N	{}
M	{N}
H	{A, C, N, M}

[0037] Accordingly, the *neighborhood-based causality* ($€_{nb}$) is defined as the average $€_{K\&M}(i)$ for all $i \in Q(j)$ as follows:

$$€_{nb}(j) = \frac{\sum_{i \in Q(j)} €_{K\&M}(i)}{|Q(j)|} \quad (10)$$

[0038] The intuition behind this metric is that accounts who are retweeting a message that was tweeted/retweeted by several causal users are potential for PSM accounts. *Weighted neighborhood-based causality* ($€_{wnb}$) is defined as follows:

$$€_{wnb}(j) = \frac{\sum_{i \in Q(j)} w_i * €_{K\&M}(i)}{\sum_{i \in Q(j)} w_i} \quad (11)$$

[0039] The intuition behind the metric $€_{wnb}$ is that the users in Q may not have the same impact on user *j* and thus different weights w_i are assigned to each user *i* with $€_{K\&M}(i)$.

Problem Statements

[0040] The goal is to find the potential PSM accounts from the cascades. Assigning a score to each user and applying threshold-based algorithm is one way of selecting users. As discussed above, causality metrics were defined where each of the metrics or combination of metrics can be a strategy for assigning scores. Users with high values for causality metrics are more likely to be PSM accounts – this relationship is demonstrated between these measurements and the real world by identifying accounts deactivated eventually.

IPR2025-01398
CrowdStrike EX1002 Page 157

Problem 1. (Threshold-Based Problem). *Given a causality metric ϵ_k where $k \in$ {K&M, rel, nb, wnb}, parameter θ, set of users U, we wish to identify set {u s.t. $\forall u \in U$, $\epsilon_k(u) \geq \theta$}.*

[0041] It was found that considering a set of cascades as a hypergraph where users of each cascade are connected to each other can better model the PSM accounts. The intuition is that densely connected users with high values for causality are the most potential PSM accounts. In other words, the interest is in selecting a user if (1.) it has a score higher than a specific threshold or (2.) it has a lower score but occurs in the cascades where high score users occur. Therefore, the *label propagation* problem is defined as follows:

Problem 2. (Label Propagation Problem). *Given a causality metric ϵ_k where $k \in$ {K&M, rel, nb, wnb}, parameters θ, λ, set of cascades $T = \{\tau_1, \tau_1, ..., \tau_n\}$, and set of users U, we wish to identify set $S : S_1, S_2, ..., S_l, ..., S_{|U|}$ where $S_l = \{u | \forall \tau \in T, \forall u \in (\tau \backslash S_{l-1}), \epsilon_k(u) \geq (H^l_\tau - \lambda)\}$ and $H^l_\tau = \{min\ \epsilon_k(u))\ s.t.\ \forall u \in \tau \wedge u \in \bigcup S_l\}_{l \in [1,l)}$.*

[0042] In another problem statement, the framework seeks to determine if causal or decay-based metrics could be leveraged for identifying PSM accounts at early stage of their life span. Note each user can now be represented by a causality vector $x \in \mathbb{R}^d$, which is computed by any of the causality metrics (summarized in Table 3 below) over a period of time. The problem of early identification of PSM accounts is formally defined as follows.

Table 3: Summary of the causality metrics.

Name	Definition	Ref.
$\epsilon_{K\&M}(i)$	Given user i, it measures $\epsilon_{K\&M}(i)$ over the time interval I.	(Anonymous 2017)
$\mathcal{E}^I_{rel}(i)$	Measures the degree of causality exhibited by user i over I while magnifying the difference between $P_{i,j}$ and $P_{v,j}$.	(Anonymous 2017)
$\mathcal{E}^I_{nb}(j)$	Measures how likely user j has adopted behavior from users in $Q(j)$ over I.	(Anonymous 2017)
$\mathcal{E}^I_{wnb}(j)$	Measures how likely user j has adopted behavior from users in $Q(j)$ with different impact on j and over I.	(Anonymous 2017)
$\mathcal{E}^I_k(i)$	Decay-based variant of $\mathcal{E}^I_k$ when $k \in \{K\&M, rel, nb, wnb\}$.	This paper

IPR2025-01398
CrowdStrike EX1002 Page 158

Problem 3 (Early Identification of PSM Accounts)

[0043] Given *an action log* **A**, *and user u where* $\exists t$ *s.t.* $(u, m, t) \in \mathbf{A}$, *and* $\nexists m, t'$ *s.t.* $(u, m, t') \in \mathbf{A} \wedge t' > t$, *our goal is to determine if u's account shall be suspended given its corresponding causality vector x computed till some time point* t'', *where* $t'' \ll t$.

ALGORITHMS

Algorithm for Threshold-based Problems

[0044] With respect to Problem 1 discussed above, the framework uses a map-reduce programming model to calculate causality metrics. In this approach, users are selected with causality value greater than or equal to a specific threshold. This approach is referred to as the *Threshold-Based Selection Approach.*

Label Propagation Algorithms

[0045] With respect to Problem 2, label propagation algorithms iteratively propagate labels of a seed set to their neighbors. All nodes or a subset of nodes in the graph are usually used as a seed set. A Label Propagation Algorithm (Algorithm 1) is proposed to solve problem 2. First, users with a causality value greater than or equal to a specific threshold (i.e. 0.9) are taken as the seed set. Then, in each iteration, every selected user u can activate user u' if the following two conditions are satisfied: (1.) u and u' have at least one cascade (action) in common and (2.) $\epsilon_k(u')) \geq \epsilon_k(u) - \lambda$, $\lambda \in (0, 1)$. Note that a minimum threshold, such as 0.7, is set so that all users are supposed to satisfy it. In this algorithm, inputs are a set of cascades (actions) T, causality metric ϵ_k and two parameters θ, λ in (0, 1). This algorithm is illustrated by a toy example:

Example 2. *Consider three cascades* {{A, B, G}, {A, B, C, D, E, G, H, I}, {E, H, I}} *as shown in hypergraph of FIGS.* ***1A-1D****.*

[0046] Assuming *the minimum acceptable value is set as 0.7; in this case, users C and E would not be activated in this algorithm. Assuming two parameters* $\theta = 0.9$, $\lambda = 0.1$, *both users* A *and* G *get activated (FIG.* ***1A****). Note that an active user is able to activate inactive ones if* (1.) *it is connected to the inactive user in the hypergraph,* (2.) *score of the inactive user meets the threshold. In the next step, only user* B *will be influenced by* G ($0.82 \geq 0.92 - 0.1$) *as it is shown in FIG.* ***1B****.*

IPR2025-01398
CrowdStrike EX1002 Page 159

Then, user D *will be influenced by user* B (0.73 ≥ 0.82 − 0.1). *In the next step (FIG.* ***1D****), the algorithm terminate since no new user is adopted. As it is shown, user* I *and* H *are not influenced although they have larger values off in comparison with user* D.

[0047] **Proposition 1.** *Given a set of cascades T, a threshold θ, parameter λ, and causality values* ϵ_k *where* $k \in \{K\&M, rel, nb, wnb\}$, *ProSel returns a set of users* $R = \{u | \epsilon_k(u) \geq \theta$ *or* $\exists u'$ *s.t.* $u', u \in T, \epsilon_k(u) \geq \epsilon_k(u') - \lambda$ *and* u' *is picked*$\}$. *Set R is equivalent to the set S in Problem 2.*

[0048] **Proposition 2.** *The time complexity of Algorithm ProSel is* $O(|T| \times avg(log(|T|)) \times |U|)$

[0049] Previous causal metrics do not take into account time-decay effect while computing causality scores. In other words, they assume a steady trend for computing causality of users. This is considered an unrealistic assumption since causal effect of users may change over time. Here, a generic decay-based metric is introduced, which essentially addresses the time-decay effect problem by assigning different time points of a given time interval, different weights, inversely proportional to their distance from *t*. In more details, this metric does the followings: (1) breaks down the given time interval into shorter time periods, using a sliding time window, (2) deploys an exponential decay function of the form $f(x) = e^{-\sigma x}$ to account for the time-decay effect, and (3) takes average between the causality values computed using the sliding time window. We formally define ξ_k^t as follows, which by varying $k \in \{K\&M, rel, nb, wnb\}$ could be used to compute each of the previous causal metrics.

$$\xi_k^t(i) = \frac{1}{|\mathbb{T}|} \sum_{t' \in \mathbb{T}} e^{-\sigma(t-t')} \times \epsilon_k^{t'}(i) \tag{13}$$

[0050] Here σ is a scaling parameter of the exponential decay function, $\mathbb{T} = \{t' | t' = t_0 + j * \delta, j \in \mathbb{N} \wedge t' \leq t - \delta\}$, δ is a small fixed amount of time, which is used as the length of the sliding time window $\Delta = [t' - \delta, t']$. In FIG. **11**, an illustration of the proposed metric is shown.

[0051] Similar to the previous observations, FIGS. **12A-12D** depict box plots of the distributions of the decay-based metrics for active and inactive users. Likewise, we use a threshold of 0.7 for all metrics except for ξ_{wnb}^t for which we used a

IPR2025-01398
CrowdStrike EX1002 Page 160

threshold of 7. Table 5 shows F1-scores for the metrics. As before, the metrics together perform the best while the next best one is ξ_{Degree}. Improvements over the causal metrics discussed before were observed.

Table 5: F1-score results for PSM accounts using each decay-based metric. The best performance is in bold.

Metric	F1-score				
	10%	20%	30%	40%	50%
ξ_{Degree}	0.44	0.46	0.47	0.5	0.53
ξ_{PR}	0.38	0.38	0.39	0.41	0.42
ξ_{Ev}	0.52	0.53	0.54	0.56	0.59
ξ_{Comb}	0.54	0.55	0.57	0.58	0.6
ALL	**0.57**	**0.59**	**0.6**	**0.62**	**0.63**

Table 6: F1-score results for PSM accounts using each decay-based metric by considering communities. The best performance is in bold.

Metric	F1-score				
	10%	20%	30%	40%	50%
ξ_{Degree}	0.49	0.51	0.52	0.54	0.57
ξ_{PR}	0.4	0.43	0.46	0.49	0.5
ξ_{Ev}	0.56	0.57	0.58	0.6	0.61
ξ_{Comb}	0.57	0.58	0.6	0.62	0.63
ALL	0.59	0.61	**0.62**	**0.65**	**0.67**

Leveraging Community Structure Aspects of Causality

[0052] Having defined the problem of early identification of PSM accounts, the question of whether leveraging community structure of PSM accounts could boost the classification performance by checking if users in a given community establish stronger causal-based relationships with each other compared to the rest of the communities is examined.

[0053] To answer this question, since network structure and its underlying graph are not available, a hypergraph **G** = (**V**, **E**) from **A** is built by connecting any pairs of users who have posted same message *chronologically*. In this hypergraph, **V** is a set of vertices (i.e. users) and **E** is a set of directed edges between users. For sake of simplicity and without loss of generality, we make the edges of this hypergraph, undirected. Next, the LOUVAIN algorithm is utilized to find the partitions **C** = $\{C_1, C_2, ..., C_k\}$ of k communities over **G**. LOUVAIN is chosen due to its fast runtime and scalability. After detecting the communities, the following two-sample t-test is performed to answer the question:

IPR2025-01398
CrowdStrike EX1002 Page 161

$$H_0 : v_a \geq v_b, \quad H_1 : v_a < v_b \tag{14}$$

where the null hypothesis is: *users in a given community establish weak causal relations with each other as opposed to the other users in other communities*. To work this out, two vectors v_a and v_b are constructed as follows. v_a is generated by computing Euclidean distances between each pair of users (u_i, u_j) who are from same community $C_1 \in \mathbf{C}$. Therefore, v_a contains exactly $\frac{1}{2}\sum_{i=1}^{|C_1|} |C_1|(|C_1|-1)$ elements in total. Likewise, v_b is constructed of size $\sum_{i=1}^{|C_1|} |C_1|$ by computing Euclidean distance between each user u_i in community $C_1 \in \mathbf{C}$, and a random user u_k chosen from the rest of the communities, i.e., $\mathbf{C} \setminus C_1$. The null hypothesis is rejected at significance level $a = 0.01$ with the p-value of 4.945e-17. Thus, it is concluded that users in same communities are more likely to establish stronger causal relationships with each other than other users in rest of the communities. The answer to the posed question is thus positive. Note for brevity, t-test results for 10% of the training set were only reported, while making similar arguments for other percentages is straight-forward.

[0054] FIGS. **13A-13D** shows box plots of the distributions of users using the decay-based metrics and the communities. A clearer distinction was observed between active and suspended accounts, using these new metrics. Again, a threshold of 7 for ξ^{λ}_{Sent} was used and 0.7 for the rest. Table 6 shows FI-scores for the metrics. The metrics together reach the best performance while the next best one is ξ^{λ}_{Sent}. The results show improvements over previous results.

Algorithm 1 Causal Community Detection-based Classification (C^2DC)

Require: Training samples $\{x_1, \ldots, x_N\}$, test samples $\{x'_1, \ldots, x'_n\}$, G, k
Ensure: Predicted labels $\{y'_1, \ldots, y'_n\}$

```
C ← LOUVAIN(G)
for each x'_i do
    C_i ← C' ∈ C s.t. x'_i ∈ C'
    D ← {}
    for each x_j ∈ C_i do
        d_ij ← ||x'_i − x_j||_2
        D ← D ∪ {d_ij}
    end for
    K ← KNN(D, k)
    y'_i ← DOMINANT-LABEL(K)
end for
```

[0055] Next, a causal community detection-based classification algorithm, namely C^2DC, is utilized that takes into account causality scores of users

IPR2025-01398
CrowdStrike EX1002 Page 162

and their underlying community structure to perform early detection of PSM accounts. First step of the algorithm involves finding the communities of the hypergraph. The LOUVAIN algorithm is used to find the partitions of the hypergraph. Then, in the second step of our algorithm, each user is classified as either PSM or not based on the available labels of her nearby peers in the same community. Specifically, to determine the label of an unseen user, we first find the community she belongs to. Then, we use the K-NEAREST NEIGHBORS (KNN) algorithm to compute her *k* nearest neighbors in the same community as hers, based on Euclidean distances between their causality vectors. Finally, the unseen user is labeled based on the majority label of her *k* nearest neighbors in the community. Note the merit of using community structure over merely using KNN is that, communities can give finer-grained and more accurate sets of neighbors sharing similar causality scores, The proposed framework is shown in Algorithm 1.

ISIS DATASET

[0056] A dataset was collected from Feb. 2016 to May 2016, consisting of ISIS related tweets/retweets in Arabic. . The dataset contains different fields including tweets and the associated information such as user ID, re-tweet ID, hashtags, content, date and time. The dataset also contains user profile information, including name, number of followers/followees, description, location, etc. About 53M tweets are collected based on the 290 hashtags such as Terrorism, State of the Islamic-Caliphate, Rebels, Burqa State, and Bashar-Assad, Ahrar Al-Sham, and Syrian Army. In this dataset, about 600K tweets have at least one URL (i.e., event) referencing one of the social media platforms or news outlets. There are about 1.4M of paired URLs which are denoted by $\mu_1 \rightarrow \mu_2$ and indicates a retweet (with the URL μ_2) of the original tweet (with the URL μ_1). In this paper, we only use tweets (more than 9M) associated with viral cascades. The statistics of the dataset are presented in Table III discussed in details below.

TABLE III: Statistics of the dataset

Name	Values
Tweets	9,092,978
Cascades	35,251
Users	1,249,293
Generator users	8,056

IPR2025-01398
CrowdStrike EX1002 Page 163

[0057] **Cascades.** The present disclosure aims to identify PSM accounts - which in this dataset are mainly social bots or terrorism-supporting accounts that participate in viral cascades. The tweets that have been retweeted from 102 to 18,892 times. This leads to more than 35k cascades which are tweeted or retweeted by more than 1M users. The distribution of the number of cascades vs cascade size is illustrated in FIG. **2A**. There are users that retweet their own tweet or retweet a post several times; however, only the first tweet/retweet of each user for a given cascade is considered. In other words, duplicate users are removed from the cascades, which make the size of the viral cascades from 20 to 18,789 as shown in FIG. **2B**. The distribution of the cascades over the cascade life span is illustrated in FIG. **3A**. Cascades took from 16 seconds to more than 94 days to complete.

[0058] **Users.** There are more than 1M users that have participated in the viral cascades. FIG. **3B** demonstrates the cumulative distribution of the number of times a user have participated in the cascades. As it is shown, the larger the support value is, the less number of users exists. Moreover, users have tweeted or retweeted posts from 1 to 3,904 times and on average each user has participated more than 7 times.

TABLE IV: Status of a subset of the users in the dataset.

Name	Active	Inactive	Total
Users	723,727	93,770	817,497
Generator users	7,243	813	8,056

[0059] **User's Current Status.** Key users are selected that have tweeted or retweeted a post in its early life span - among first half of the users (according to Definition 2, $\phi = 0.5$), and check whether they are active or not. Accounts are not active if they are suspended or deleted. More than 88% of the users are active as shown in Table IV. The statistics of the generator users are also reported. Generator users are those that have initiated a viral cascade. As shown, 90% of the generator users are active as well. Moreover, there are a significant number of cascades with hundreds of inactive users. The number of inactive users in every cascade is illustrated in FIG. **3C**. Inactive users are representative of automatic and terrorism accounts aiming to disseminate their propaganda and manipulate the statistics of the hashtags of their interest.

IPR2025-01398
CrowdStrike EX1002 Page 164

[0060] **Generator Users.** In this part, users that have generated the viral tweets are only considered. According to Table IV, there are more than 7k active and 800 inactive generator users. That is, more than 10% of the generator users are suspended or deleted, which means they are potentially automated accounts. The distribution of the number of tweets generated by generator users shows that most of the users (no matter active and inactive) have generated a few posts (less than or equal to 3) while only a limited number of users are with a large number of tweets..

[0061] **Impact of μ_1 on μ_2.** In this part, the impact of the URL μ_1 on μ_2 is investigated. Accordingly, the dataset contains 35K cascades (i.e., sequences of events) of different sizes and durations, some of which contain paired URLs in the aforementioned form. After pre-processing and removing duplicate users from cascades (those who retweet themselves multiple times), cascades sizes (i.e. number of associated postings) vary between 20 to 9,571 and take from 10 seconds to 95 days to finish. The log-log distribution of cascades vs. cascade size and the cumulative distribution of cascades are depicted in FIG. 5.

The statistics of the dataset are presented in Table V. For labeling, the Twitter API is checked to examine whether the users have been suspended (labeled as PSM) or they are still active (labeled as normal). According to Table V, 11% of the users in the dataset are PSMs and others are normal. The total number of PSM accounts that have been suspended by Twitter is depicted in each cascade in FIG. 6. Finally, it should be noted that these accounts mostly get suspended manually by Twitter based on reports the platform receives from its own users.

Table V: Description of the dataset

Name	Value	
# of Cascades	35K	
# of Tweets/Retweets	2.8M	
	PSM	Normal
# of Users	64,484	536,609
# of Single URLs	104,948	536,046
# of Paired URLs	200,892	1,123,434

Social Media Platforms and News Outlets

[0062] Twitter deploys a URL shortener technique to leave more space

IPR2025-01398
CrowdStrike EX1002 Page 165

for content and protect users from malicious sites. To obtain the original URLs, a URL unshortening tool is used. This tool obtains the original links contained in the tweets of the dataset.

[0063] A number of major and well known social media platforms are considered, including Facebook, Instagram, Google, and YouTube. About the dichotomy of mainstream and alternative media, it is notable to mention that most criteria for determining whether a news source counts as either of them are based on a number of factors including but not limited to the content and whether or not it is corporate owned. However, a key difference between these two sources of media comes from the fact that all of mainstream media is profit oriented, in contrast to alternative media. It is further noted that, for the most part, mainstream media is considered a more credible source than alternative media, although the reputation has been recently tainted by the fake news.

[0064] Popular news outlets such as The New York Times and The Wall Street Journal are considered as mainstream and less popular ones as alternatives. In Table VI, the total number of paired URLs (i.e. $\mu_1 \rightarrow \mu_2$) in which the original URL (i.e. μ_1) corresponds to each social media platform with at least one event in the dataset is summarized. In Table VII, the total number of paired URLs whose original URL belongs to the mainstream and alternative news sources is summarized. Table VIII shows the breakdown of the number of paired URLs for the PSM and normal users. Table IX further demonstrates some examples of the mainstream and alternative news URLs occurrence used in the work.

Table VII: News sources' total paired URLs ($\mu_1 \rightarrow \mu_2$) with at least one event in the dataset for the PSM and normal users.

Platform	PSM	Normal
Twitter	139,940	918,803
Facebook	878	4,017
Instagram	0	2,857
Google	163	132
Youtube	24,724	72,890

Table VIII: Total number of paired URLs of the form $\mu_1 \rightarrow \mu_2$ with at least one event in the dataset for the PSM/normal users and for all platforms.

IPR2025-01398
CrowdStrike EX1002 Page 166

News Source	PSM	Normal
Mainstream	0	286
Alternatives	35,187	124,449

Table: IX: Examples of mainstream and alternative news used in this work. Popular news outlets are assigned to mainstream and the rest to alternatives.

	⇢ Twitter	⇢ Facebook	⇢ Instagram	⇢ Google	⇢ Youtube	⇢ Mainstream	⇢ Alternatives
Twitter ⇢	109,354/766,617	568/3,843	229/2,461	120/582	11,992/59,889	90/688	17,557/84,925
Facebook ⇢	655/3,108	4/41	3/9	2/1	87/281	0/1	127/576
Instagram ⇢	0/2,362	0/11	0/25	0/2	0/161	0/2	0/294
Google ⇢	134/74	0/0	0/1	0/0	12/53	0/0	17/4
Youtube ⇢	14,004/56,545	132/312	23/211	22/32	6,799/7,529	13/48	3,731/8,213
Mainstream ⇢	0/189	0/1	0/1	0/0	0/13	0/1	0/81
Alternatives ⇢	21,047/95,641	145/767	45/318	59/64	3,862/9,199	26/122	10,003/18,338

Table VI: Social media platform's total number of paired URLs of the form $\mu_1 \dashrightarrow \mu_2$ with at least one event in the dataset for the PSM and normal users

Mainstream	Alternatives
https://www.nytimes.com	https://www.rt.com
https://www.reuters.com	https://www.arabi21.com
https://www.wsj.com	https://www.7adramout.net
https://www.nbcnews.com	https://www.addiyar.com
https://www.ft.com	https://zamnpress.com

Temporal Analysis

[0065] Here, the differences between PSM accounts with their counterparts, normal users, through temporal analysis of their posted URL's are presented.

[0066] In FIG. **7**, the daily occurrence of the paired URL's over the span of 43 days for both PSM and normal users is depicted. The dataset has a larger number of normal users and higher number of the posted URL's compared to PSM accounts. Therefore, it is reasonable to observe more activity from normal users than PSMs. For both groups of users, a similar trend in spikes and their duration is observed. As it is seen, distinguishing between PSMs and normal users merely based on the occurrence of URLs and their patterns is not reliable. Therefore, experiments are conducted using a more sophisticated statistical technique known as "Hawkes Process".

Framework

[0067] In the previous section, a data analysis to demonstrate differences between PSM accounts and normal users in terms of URLs they usually

IPR2025-01398
CrowdStrike EX1002 Page 167

post on Twitter was presented. The next step is to assess their impact via a well known mathematical process called "Hawkes Process". A Hawkes process is used since the platforms introduced are dependent and thus not disjoint, meaning that they are affected by each other, while they have their own background events.

A. Hawkes Process

[0068] In many scenarios, one has to deal with timestamped events such as the activity of users on a social network recorded in continuous time. An important task then is to estimate the influence of the nodes based on their timestamp patterns. Point process is a principled framework for modeling such event data, where the dynamic of the point process can be captured by its conditional intensity function as follows:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{E}(N(t+\Delta t) - N(t) | \mathcal{H}_t)}{\Delta t} = \frac{\mathbb{E}(dN(t) | \mathcal{H}_t)}{dt}$$

[0069] Where $\mathbb{E}(dN(t)|\mathcal{H}_t)$ is the expectation of the number of events which happened in the interval $(t, t+dt)$ given the historical observations $\mathcal{H}_t$, and $N(t)$ records the number of events before time t. Point process can be equivalently represented as a counting process $N = \{N(t) | t \in [0, T]\}$ over the time interval $[0, T]$.

The Hawkes process framework has been used in many problems that require modeling complicated event sequences where historical events have impact on future ones. Examples include but are not limited to financial analysis, seismic analysis, and social network modeling. One-dimensional Hawkes process is a point process N_t with the following particular form of intensity function:

$$\lambda(t) = \mu + \alpha \int_{-\infty}^{t} g(t-s) dN_s = \mu + \alpha \sum_{t_i < t} g(t - t_i)$$

[0070] where $\mu > 0$ is the exogenous base intensity (i.e. the background rate of events), t_i are the time of events in the point process before time t, and $g(t)$ is the decay kernel.

[0071] Here, the exponential kernal of the form $g(t) = we^{-wt}$ is used, but adapting to the other positive terms is straightforward. The second part of the

IPR2025-01398
CrowdStrike EX1002 Page 168

above formulation captures the self exciting nature of the point processes – the occurrence of events in the past has a positive impact on future ones. Given a sequence of events $\{t_i\}_{i=1}^{n}$ observed in [0, *T*] and generated from the above intensity function, the log-likelihood function can be obtained as follows:

$$\mathcal{L} = \log \frac{\prod_{i=1}^{n} \lambda(t_i)}{\exp \int_0^T \lambda(t) dt} = \sum_{i=1}^{n} \log \lambda(t_i) - \int_0^T \lambda(t) dt$$

Here, the focus is on multi-dimensional Hawkes processes which is defined by *U* – dimensional point process $N_t^u, u = 1, \ldots, U$. In other words, *U* Hawkes processes are coupled with each other – each Hawkes process corresponds to one of the platforms and the influence between them is modeled using the mutually exciting property of the multi-dimensional Hawkes processes. The following formulation to model the influence of different events on each other is formally defined:

$$\lambda_u(t) = \mu_u + \sum_{i:t_i<t} a_{u u_i} g(t - t_i)$$

[0072] where $\mu_u \geq 0$ is the base intensity for the *u*-th Hawkes process. The coefficient $a_{u u_i} \geq 0$ captures the mutually-exciting property between the *u*-th and u_i-th dimension are more likely to trigger an event in *u*-th dimension in the future. More intuitively, an event on one point process can cause an impulse response on other processes, which increases the probability of an event occurring above the processes' background rates. To reiterate, each URL is attributed to an event, i.e. if the URL u_1 triggers the URL u_2 (i.e. $u_1 \rightarrow u_2$), then $a_{u_2 u_1} \geq 0$.

[0073] In FIG. **8**, a multivariate example of three different streams of events, e_0, e_1, and e_2. As illustrated, e_0 is caused by the background rate $\lambda(t)_0$ and has an incluence on itself and e_1. On the other hand, e_1 is caused by $\lambda(t)_1$ and has an infuence on e_2. Simply put, a background event in e_0 induces impulse on responses on processes e_1 and e_2. Accordingly, the caused child event in e_1 leads to another child event in e_2.

[0074] Consider an infectivity matrix $\boldsymbol{A} = [a_{u u_i}] \in \mathbb{R}^{D \times D}$ which collects

IPR2025-01398
CrowdStrike EX1002 Page 169

the self triggering coefficients between Hawkes processes, and $D = 7$ is the number of processes (i.e. platforms). Each entry in this matrx indicates the strength of influence each platform has on other platforms.

B. Methodology

[0075] The goal is to assess the influence of the PSM accounts in the dataset via their posted URLs. The URLs posted by two groups of users are considered: (1) PSM accounts and (2) normal users. For both groups, a Hawkes model with $K = 7$ point processes each is fitted for each category of social media and news outlets.

[0076] In each of the Hawkes models, every process is able to influence all the others including itself, which allow for the estimation of the strength of connections between each of the seven categories for both groups of users, in terms of how likely an event (i.e. the posted URL) can cause subsequent events in each of the groups.

The ADM4algorithm is used and the methodology presented for fitting the Hawkes processes for both PSM and normal users is used. ADM4is an efficient optimization that estimates the parameters $\boldsymbol{A}$ and μ by maximizing the regularized log-likelihood $\mathcal{L}(\boldsymbol{A}, \mu)$:

$$\min_{\boldsymbol{A}\geq 0, \mu \geq 0} -\mathcal{L}(\boldsymbol{A}, \mu) + \lambda_1 \|\boldsymbol{A}\|_* + \lambda_2 \|\boldsymbol{A}\|_1$$

[0077] where $\mathcal{L}(\boldsymbol{A}, \mu)$ can be obtained by substituting ___ from Equation ___ into Equation ___. Also, $\|\boldsymbol{A}\|_*$ is the nuclear form of matrix $\boldsymbol{A}$, and is defined as the sum of its singular value.

[0078] Two different sets of URLs posted by the PSM account and normal users are considered by selected URLs that have at least one event in Twitter (i.e. posted by a user). For each group, a matrix $\boldsymbol{W} \in \mathbb{N}^{T \times D}$ with D =7, whose entries are sequences of events (i.e. posted URLs) observed during time period T are constructed. Each sequence of events is of the form $S = \{(t_i, u_i)\}_{i=1}^{n_i}$where n_i is the number of events occurring at the *ui*-th dimension (i.e., URLs posted containing one of the 7 platforms).

C. Settings

IPR2025-01398
CrowdStrike EX1002 Page 170

[0079] The ADM4 algorithm is adopted, which implements parametric interference for Hawkes processes with an exponential kernel and a mix of Lasso and nuclear regularization. The infectivity matrix A, base intensities μ,and decays $\beta \in \mathbb{R}$ are initialized randomly.

[0080] The number of nodes is set to D=7 to reflect the 7 platforms used in this analysis. Level of penalization is set to C = 1000, and the ratio of Lasso-Nuclear regularization mixing parameter is set to 0.5. Finally, he maximum number of iterations for solving the optimization is set to 50 and the tolerance of solving algorithm is set to 1e – 5.

CAUSALITY ANALYSIS

[0081] The behavior of the causality metrics is examined. Users are analyzed based their current account status in Twitter. A user is labeled as active (inactive) if the account is still active (suspended or deleted).

[0082] **Kleinberg-Mishra Causality.** Users were studied that had a causality value of $\epsilon_{K\&M}$ greater than or equal to 0.5. As expected, inactive users exhibit different distribution from active users (FIG. **4A**). It was noted that significant differences are present - more than 75% of the active users are distributed between 0.5 and 0.62, while more than 50% of the inactive users are distributed from 0.75 to 1. Also, inactive users have larger values of mean and median than active ones. Note that number of active and inactive users is 404,536 and 52,452. This confirms that this metric is a good indicator to discriminate PSM users from the normal users. Additionally, several related causality measures and a label-propagation algorithm for finding "slave PSM accounts" have been developed. **Relative Likelihood Causality.** This metric magnifies the interval between every pairs of the probabilities that measures the causality of the users; therefore, the values vary in a wide range. FIG. **4B** displays the distribution of users having relative likelihood causality of greater than or equal to two. In this metric, 1,274 users get very large values. For the sake of readability, very large values are replaced with 34.0 in FIG. **4B**. More than 50% of the inactive users get values greater than 32, while the median of active users is 2.48. More than 75% of the active users are distributed in the range of (2, 4). Note that number of active and inactive users in this figure is 3,563 and 1,041, respectively. That is, using this metric and filtering users with the relative likelihood

IPR2025-01398
CrowdStrike EX1002 Page 171

greater than a threshold, leads to the good precision. For example, the threshold in FIG. **4B** is set to 2 - the precision is more than 0.22 for inactive class. Considering users with a very large value can lead to a precision of more than 0.5 and uncovering a significant number of PSMs - 638 inactive users.

[0083] **Neighborhood-Based Causality.** The users having a causality value of ϵ_{nb} greater than or equal to 0.5 were studied. As expected, inactive users exhibit different distribution from active users as shown in FIG. **4C**. Also, inactive users are mostly distributed in the higher range and have larger values of mean and median than active ones. More than 75% of the active users are distributed between 0.5 and 0.6, while more than 50% of the inactive users are distributed from 0.6 to 1. Therefore, increasing the threshold results in the higher precision for the PSM users. Note that the number of active and inactive users is 85,864 and 10,165.

[0084] **Weighted Neighborhood-Based Causality.** This metric is the weighted version of the previous metric (ϵ_{nb}). A weight is assigned to each user in proportion to her participation rate in the viral cascades. FIG. **4D** shows the distribution of users with ϵ_{wnb} greater than or equal to 0.5. This metric also displays different distribution for active and inactive users. More than 75% of the active users are distributed between 0.5 and 0.6, while more than 50% of the inactive users are distributed from 0.6 to 1. Note that the number of active and inactive users of ϵ_{wnb} is 52,346 and 16,412. In other words, this metric achieves the largest precision compared to other metrics, 0.24. Clearly, increasing the threshold results in the higher precision for the PSMs.

RESULTS AND DISCUSSION

[0085] The Scala Spark and Python 2.7x were run on an Intel Xeon CPU (1.6 GHz) with 128 GB of RAM running Windows 7. The parameter ϕ was set to label key users 0.5 (Definition 2). Thus, for the users that participate in the action before the number of participants gets twice.

[0086] In the following sections, two proposed approaches are considered: (1) *Threshold-Based Selection Approach* - selecting users based on a specific threshold, (2) *Label Propagation Selection Approach* - selecting by applying Algorithm 1. The intuition behind this approach is to select a user if it has a score higher than a threshold or has a lower score but occurs in the cascades that high

IPR2025-01398
CrowdStrike EX1002 Page 172

score users exist. The methods are evaluated based on true positive (True Pos), false positive (False Pos), precision, the average (Avg CS) and median (Med CS) of cascade size of the detected PSM accounts. Note that in our problem, precision is the most important metric. The main reason is labeling an account as PSM means it should be deleted. However, removing a real user is costly. Therefore, it is important to have a high precision to prevent removing real user.

EXISTING METHOD

[0087] In the proposed method, all features that could be extracted from the dataset include tweet syntax (average number of hashtags, average number of user mentions, average number of links, average number of special characters), tweet semantics (LDA topics), and user behavior (tweet spread, tweet frequency, tweet repeats). Three existing methods were to detect PSM accounts: 1) *Random* selection: This method achieves the precision of 0.11. This also presents that the data is imbalanced and less than 12% of the users are PSM accounts. 2) *Sentimetrix* (*Sentimet.*): The data was clustered by DBSCAN algorithm. The labels were then propagated from 40 initial users to the users in each cluster based on the similarity metric. Support Vector Machines (SVM) were used to classify the remaining PSM accounts [10]. 3) *Classification* methods: In this experiment, the same labeled accounts were used as the previous experiment and different machine learning algorithms were applied to predict the label of other samples. Features based on the limitations of access to data into three categories were then grouped together. First, only content information (*Content*) is used to detect the PSM accounts. Second, content independent features (*NoCont.*) is used to classify users. Third, all features (*All feat.*) are applied to discriminate PSM accounts. The best result for each setting is when we apply Random Forest using all features. According to the results, this method achieves the highest precision of 0.16. Note that, most of the features used in the previous work and our baseline take advantage of both content and network structure. However, there are situations that the network information and content do not exist. In this situation, the best baseline has the precision of 0.15. The average (Avg CS) and median (Med CS) of the size of the cascades were studied in which the selected PSM accounts have participated. Table X also illustrates the false positive, true positive and precision of different methods.

TABLE X: Existing Methods – number of selected users as PSM

IPR2025-01398
CrowdStrike EX1002 Page 173

Method	False Pos	True Pos	Precision	Avg CS	Med CS
Random	80,700	10,346	0.11	289.99	184
Sentimet.	640,552	77,984	0.11	261.37	171
Content	292,039	43,483	0.13	267.66	174
NoCont.	357,027	63,025	0.15	262.97	172
Allfeat.	164,012	31,131	**0.16**	273.21	176

THRESHOLD-BASED SELECTION APPROACH

[0088] In this experiment, all the users were selected that satisfied the thresholds and checked whether they were active or not. A user is *inactive* if the account is suspended or closed. Since the dataset is not labeled, inactive users were labeled as PSM accounts. The threshold for all metrics was set to 0.7 except for relative likelihood causality (ϵ_{rel}), which is set to 7. Two types of experiments were conducted: first, user selection for a given causality metric was studied. This approach was further studied using the combinations of metrics.

[0089] **Single Metric Selection.** In this experiment, users were selected based on each individual metric. As expected, these metrics can assist in filtering a significant amount of active users and overcome the data imbalance issue. Metric $\epsilon_{K\&M}$ achieves the largest recall in comparison with other metrics. However, it has a maximum number of false positives. Table XI shows the performance of each metric. The precision value varies from 0.43 to 0.66 and metric ϵ_{wnb} achieves the best value. Metric ϵ_{rel} finds the more important PSM accounts with average cascade size of 567.78 and median of 211. In general, the detected PSM accounts have participated in the larger cascades in comparison with baseline methods. It was also observed that these metrics cover different regions of the search area. In other words, they select different user sets with little overlap between each other. The common users between any two pairs of the features are illustrated in Table XII. Considering the union of all metrics, 36,983 and 30,353 active and inactive users are selected, respectively.

TABLE XI: Threshold-based selection approach - number of selected users using single metric

Method	False Pos	True Pos	Precision	Avg CS	Med CS
Allfeat.	164,012	31,131	**0.16**	273.21	176
NoCont.	357,027	63,025	0.15	262.97	172
$\epsilon_{K\&M}$	36,159	27,192	0.43	383.99	178
ϵ_{rel}	693	641	0.48	**567.78**	**211**

IPR2025-01398
CrowdStrike EX1002 Page 174

ϵ_{nb}	2,268	2,927	0.56	369.46	183.5
ϵ_{wnb}	7,463	14,409	**0.66**	311.84	164

TABLE XII: Threshold-based selection approach - number of common selected users using single metric

Status	**Active**			**Inactive**		
Method	ϵ_{rel}	ϵ_{nb}	ϵ_{wnb}	ϵ_{rel}	ϵ_{nb}	ϵ_{wnb}
$\epsilon_{K\&M}$	404	1,903	6,992	338	2,340	11,748
ϵ_{rel}		231	175		248	229
ϵ_{wnb}			1,358			1,911

[0090] **Combination of Metrics Selection.** According to Table XII, most of the metric pairs have more inactive users in common than active users. The experiment attempted to select users that satisfy the threshold for at least three metrics. It was determined that 1,636 inactive users were obtained out of 2,887 selected ones, which works better than $\epsilon_{K\&M}$ and ϵ_{rel} while worse than ϵ_{nb} and ϵ_{wnb}. In brief, this approach achieves precision of 0.57. Moreover, the number of false positives (1,251) was lower than most of the other metrics.

LABEL PROPAGATION SELECTION APPROACH

[0091] In label propagation selection, a set of users were selected that have a high causality score as seeds, then ProSel selects users that occur with those seeds and have a score higher than a threshold iteratively. Also, the seed set in each iteration is the selected users of the previous iteration. The intuition behind this approach is to select a user if it has a score higher than a threshold or has a lower score but occurs in the cascades that high score users occur. The parameters of *ProSel Algorithm* are set as follows: $\lambda = 0.1$, $\theta = 0.9$, except for relative likelihood causality, where we set $\lambda = 1$, $\theta = 9$. Table XIII shows the performance of each metric. Precision of these metrics varies from 0.47 to 0.75 and ϵ_{wnb} achieves the highest precision. Metrics ϵ_{rel} with average cascade size of 612.04 and ϵ_{nb} with median of 230 find the more important PSM accounts. Moreover, detected PSM accounts have participated in the larger cascades compared with threshold-based selection. This approach also produces much lower number of false positives compared to threshold-based selection. The comparison between this approach and threshold-based selection is illustrated in FIG. **9**. From the precision perspective, label propagation method outperforms the threshold-based one.

IPR2025-01398
CrowdStrike EX1002 Page 175

TABLE XIII: Label propagation selection approach - number of selected users

Method	False Pos	True Pos	Precision	Avg CS	Med CS
Allfeat.	164,012	31,131	0.16	273.21	176
NoCont.	357,027	63,025	0.15	262.97	172
$\epsilon_{K\&M}$	9,305	14,176	0.60	390.52	179
ϵ_{rel}	561	498	0.47	612.04	216
ϵ_{nb}	1,101	1,768	0.62	403.55	230
ϵ_{wnb}	1,318	4,000	0.75	355.24	183.5

[0092] The number of common users selected by any pair of two metrics is also illustrated in Table XIV. It shows that our metrics are covering different regions of the search area and identifying different sets of users. In total, 10,254 distinct active users and 16,096 inactive ones are selected.

TABLE XIV: Label propagation selection approach - number of common selected users

Status	Active			Inactive		
Method	ϵ_{rel}	ϵ_{nb}	ϵ_{wnb}	ϵ_{rel}	ϵ_{nb}	ϵ_{wnb}
$\epsilon_{K\&M}$	289	581	1,122	168	1,019	2,788
ϵ_{rel}		15	6		180	102
ϵ_{wnb}			151			833

Hawkes Process Approach

[0093] The infectivity matrix for both PSM and normal users is estimated by fitting the Hawkes model described earlier. In this study, this matrix characterizes the strength of the connections between platforms and news sources. More specifically, each weight value represents the connection strength from one platform to another. In other words, each entry in this matrix can be interpreted as the expected number of subsequent events that will occur on the second group after each event on the first.

[0094] In FIG. **10** the estimated weights for all paired URLs for both PSM and normal users are depicted. The weights in both plots show that greater weights belong to processes that have impact on Twitter, i.e. "→*Twitter*". This implies

IPR2025-01398
CrowdStrike EX1002 Page 176

that both of the groups in the Twitter dataset often post URLs that ultimately have greater impact on Twitter.

[0095] Overall, the following observations can be made:

- URLs references all platforms and posted by the PSMs and regular users mostly trigger URLs that contain the Twitter domain.
- Among all platforms studied here, URLs shared from facebook.com and alternative news media contributed the most to the dissemination of malicious information from PSM accounts. In other words, they had the largest impact on making a message viral and causing the subsequent events.
- Posts that were tweeted by the PSM accounts and contained URLs from facebook.com demonstrated more influence on the subsequent retweets containing URLs from youtube.com, in contrary to the other way around. This means that ultimately tweets with URLs from Facebook will likely end up inducing external impulse on youtube.com. In contrast, URLs posted by the normal users have nearly the same impact on subsequent events regardless of the social media or news outlet used.

[0096] The above mentioned observations demonstrate the effectiveness of leveraging Hawkes process to quantify the impact of URLs posted by PSMs and regular users on the dissemination of content on Twitter. The observations we make here show that PSM accounts and regular users behave differently in terms of the URLs they post on Twitter, in that they have different tastes while disseminating URL links. Accordingly their impact on the subsequent events significantly differ from each other.

ADDITIONAL EXPERIMENTS

[0097] Additional experiments were conducted to gauge the effectiveness and efficiency (timeliness) of the above approaches that use the causality metrics as features in (1) supervised setting, and (2) for computing proximity scores between users in community detection-based framework.

[0098] Experimental Settings

[0099] To assess the effectiveness and efficiency of the metrics and the proposed framework in early detection of PSM accounts, we use different subsets of size x% of the entire time-line (from Feb 22, 2016 to May 27, 2016) of the action log **A**, by varying x as in {10, 20, 30, 40, 50}. Next, for each subset and each user i in the subset, we compute feature vector $\mathbf{x}_i$ of the corresponding 4 causality scores. The feature vectors are then fed into supervised classifiers and the proposed community detection-based algorithm to perform classification. For the sake of fair comparison,

IPR2025-01398
CrowdStrike EX1002 Page 177

both standard and decay-based metrics were performed. For both metrics, we empirically found that $p = 0.1$ and $a = 0.001$ work well. For the decay-based causality metric a sliding window of size of 5 days (i.e. $\delta = 5$) and set $\sigma = 0.001$ was assumed which were found to work well in the experiments. Note that only results for PSM (suspended) accounts were presented. Among many other supervised classifiers such as ADABOOST, LOGISTIC REGRESSION and SUPPORT VECTOR MACHINES (SVM), RANDOM FOREST (RF) with 200 estimators and 'entropy' criterion, achieved the best performance. Therefore, for brevity only results when RF is used as the classifier was reported. This aligns well with previous results on related problems such as the identification of viral cascades.

[00100] The results for the proposed community detection-based framework and both causal and decay-based metrics are presented herein. As stated earlier, for community detection the LOUVAIN algorithm was implemented since it was fast, scaled well and it did not require any parameter tuning. For the second part of the proposed algorithm, i.e., computing k nearest neighbors of each instance and the final labeling procedure, we set k = 10 as it was found to work well for our problem. It must be stressed that using KNN alone does not yield a good performance by reporting the results of KNN trained on the decay-based causality features.

[00101] For the sake of fair comparison, all approaches were implemented and run in Python 2.7x, using scikit-learn package. For the Louvain algorithm, we used the Python community detection package was used. For any approaches that required special tuning of parameters, a grid search was conducted to choose the best set of parameters.

Baseline Methods

[00102] To validate the performance of the proposed metrics and framework, the metrics and framework were compared against the following state-of-the-art baselines:

[00103] **CAUSAL (Anonymous 2017).** The proposed decay-based metrics was benchmarked against causal metrics (Anonymous 2017) via both supervised and community detection-based settings.

[00104] **SentiMetrix-DBSCAN (Subrahmanian et al. 2016).** This method was the winner of the DARPA Twitter bot challenge contest (Subrahmanian et al.

IPR2025-01398
CrowdStrike EX1002 Page 178

2016). This baseline deploys several features including tweet syntax (average number of hashtags, average number of user mentions, average number of links, average number of special characters), tweet semantics (LDA topics), and user behavior (tweet spread, tweet frequency, tweet repeats). For this baseline, a 10-fold cross validation was performed and use a held-out test set for evaluating the baseline. This baseline then uses a seed set of 100 active and 100 inactive accounts, and then use DBSCAN clustering algorithm to find the associated clusters of users. The available labels are then propagated to nearby unlabeled users in each cluster based on the Euclidean distance metric, and the labels of the remaining accounts are predicted using SVM.

[00105] **SentiMetrix-RF**. This is a variant of the previous baseline where we excluded the DBSCAN part and instead trained RF classifier using only the above features to evaluate the feature set.

Identification of PSM Accounts

[00106] For evaluating how *accurate* different approaches are in identifying PSM accounts, we use different metrics including precision, recall and F-1 measure and area under curve (AUC) for the PSM accounts. Precision, recall, FI-score and AUC results for each method are shown in FIGS. **14A-14B** and **15A-15B**. Note for each subset, independent from other subsets, a 10-fold cross validation was performed. The following observations were made.

- Community detection-based framework achieved the best performance in terms of several metrics used. This aligns well with the *t*-test results discussed earlier, meaning that taking into account community structure of PSM accounts can boost the performance.
- For the most part, the causal and decay-based metrics achieved higher classification performance than other approaches via both settings.
- Decay-based metrics are effective at identifying PSM accounts at different time points via supervised and community detection-based settings. It is posited that this lies at the inherent difference between decay-based and causal metrics– decay-based metrics take into account time-decay effect while computing the causality scores.
- Although SENTIMETRIX-DBSCAN and SENTIMETRIX-RF take advantage of several features, they were unable to defeat our approach. It is noted that most of the

IPR2025-01398
CrowdStrike EX1002 Page 179

features used in these baselines are based on content of tweets which is not always readily available.

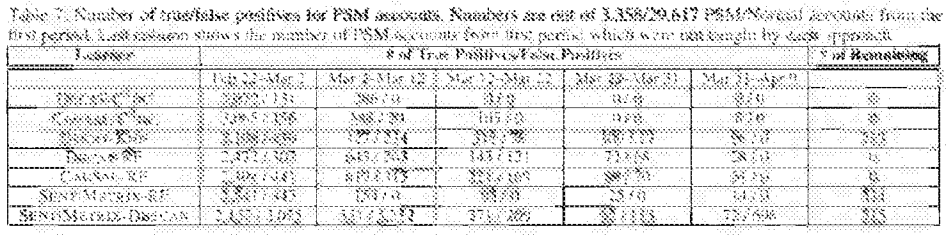
Timeliness of PSM Accounts Identification

[00107] The following experiment was conducted to gauge the ability of different approaches in identifying PSM accounts around their activity. For each approach, it was determined *how many* of the PSM accounts were active in the first 10 days of the dataset, were correctly classified (i.e., true positives) over time. Also, the framework needed to keep track of false positives to ensure a given approach does not merely label each instance as positive – otherwise a trivial approach that always would label each instance as PSM would achieve the highest performance. In addition, to figure how many days were needed to find these PSM accounts each classifier was trained using 50% of the first portion of the dataset. A held-out set was used for the rest for evaluation. Next, the misclassified PSM accounts were passed on to the next portions to see how many of them can be captured over time. The process was repeated until reaching 50% of the action log -- each time the training set was increased by adding new instances from each portion.

[00108] There are 14,841 users in the first subset from which 3,358 users are PSM. Table XV below shows the number of users from the first portion that (1) are correctly classified as PSM (out of 3,358), (2) are incorrectly classified as PSM (out of 29,617), over time. According to this table, community detection-based approaches were able to detect all PSM accounts who were active in the first 10 days of our dataset, in about a month since their first activity. In particular, DECAY-C^2DC, identified all of these PSM accounts in about 20 days since the first time they posted a message. Also, both causal and decay-based metrics when fed to RF classifier, identified all of the PSM accounts in the first period. SENTIMETRIX-DBSCAN and SENTIMETRIX-RF failed to detect all PSM accounts from the first portion, around the first 50 days of their activity. Furthermore, these two baselines generated much higher rates of false positives compared to all of the causality-based approaches.

IPR2025-01398
CrowdStrike EX1002 Page 180

Table XV:



[00109] The observations that were made aligned well with previous observations; suggesting that the community detection-based framework yields higher classification performance than other frameworks. Finally, it should be noted that to make the computations and discussion less complicated, a period of 10 days was used, therefore, the exact number of days that should pass were not known, although this will still give a good approximation of the efficiency of the given approach in early identification of PSM accounts.

Blind Validation Tests for PSM Account Detection

[00110] Blind validation tests were conducted by a third party using real social media data that the third party used for other products and services. In the test, the third party provided data without ground truth, for which results were provided in return. The third party provided performance metrics compared with an existing approach for identifying social bots. It should be noted that the evaluated version of the PSM Account Detection software was the initial version. The version disclosed has shown improved performance.

Table XVII: Results of the blind validation tests

	PSM Account Detection	Existing approach from third party
Accounts identified as social bot, suspended or offline	51	10
False positives identified as “self-promoters”	6	(not reported)
Other false positives	10	(not reported)

[00111] Further comparisons with existing methods were made to to evaluate for true positives and false positives over time. FIGS. **16A** and **16B** show

IPR2025-01398
CrowdStrike EX1002 Page 181

cumulative true positives and false positives based on the available training data. True positives (FIG. **16A**) are more rapidly identified by the PSM Account Detection compared to existing approaches, and is the only approach to identify all pathogenic accounts. Likewise, it also maintains the lowest false positive rate by a significant margin (FIG. **16B**).

[00112] Finally, as discussed, existing approaches for detecting socialbots and other malicious actors on social media are highly dependent upon content – which leads to a dramatic increase in disk space and RAM requirements. FIG. **17** compares diskspace requirements from existing approaches due to content storage as opposed to requirements to this invention, which primarily uses action logs. The substantial reduction in storage requirements enables a significant reduction in computing machinery overhead enabling the approach to be scaled in order to detect a larger number of malicious actors on social media. The reductions are by an order of magnitude and highly significant by any standard statistical measure.

[00113] Turning to FIG. **18**, an example network environment **100** for generating the social media data ultimately utilized to detect PSM accounts as described herein is shown. The network environment **100** may include a plurality of user devices **102**, and a plurality of social media computing devices, designated **104A** and **104B**. User devices **102** may be associated with users of social media located throughout the world, and the social media computing devices **104A-104B** may be associated with social media providers such as e.g. Twitter, Facebook, news websites, and the like. The user devices **102** may request and submit content to the social media computing devices **104A-104B** using a network **106** (e.g. the Internet) as indicated. Specifically, the user devices **102** may adopt various messages as defined herein hosted or otherwise provided by the social media computing devices **104A-104B**. For example, any one of the user devices **102** may log on to a social media application hosted by the social media application server **104A**, and broadcast or re-broadcast a message such as a video, text message, audio file, or the like.

[00114] The user devices **102** may be generally any form of computing device capable of interacting with the network **106** to access a social media website, such as a mobile device, a personal computer, a laptop, a tablet, a workstation, a smartphone, or other Internet- communicable device. The social media computing devices **104A-104B** may include computing devices configured for providing social

IPR2025-01398
CrowdStrike EX1002 Page 182

media websites or aspects thereof and/or any applications or websites where a user may adopt a social media message, and may be implemented in the form of servers, workstations, terminals, mainframes, storage devices, or the like.

[00115] As further shown, a computing device **108** may be communicably coupled to the network **106**. The computing device **108** may be implemented to provide the functionality described herein, and may include a laptop, desktop computer, server, or other such computing device. Any one of the social media computing devices **104A-B** or the computing device **108** may include at least one server hosting a website or providing an application. Such a server (not shown) may be a single server, a plurality of servers with each such server being a physical server or a virtual machine, or a collection of both physical servers and virtual machines. In another implementation, a cloud hosts one or more components of the network environment **100**. In this implementation, the user devices **102**, any servers, and other resources connected to the network **106** may access one or more servers to access websites, applications, web services, interfaces, storage devices, computing devices, or the like.

[00116] Referring to FIG. **19**, with reference to the process flow **300**, further details regarding the implementation of the functionality described herein relative to the network environment **100** are illustrated. In the system diagram **200**, a computing device **202** is shown which may be similar to the computing device **108** of the network environment **100** such that the computing device **202** may access user devices **102** and social media computing devices **104A-B**. In particular, the computing device **202** is configured to access social media data in the form of activity logs **206A-206C** from one or more social media data sources **204A-204C**. The computing device **202** may access such information via e.g. the network **106**, or locally, and may access the data directly from social media providers (e.g. the social media computing devices **104A-B**) or third party providers of such information. In other words, the social media data sources **204A-204C** may include databases or storage devices associated with the social media computing devices **104A-B**, or may include third party storage repositories. For example, the social media data source **204A** may include Twitter data purchased through a third party, or, the social media data source **204A** may include a storage device associated with Twitter directly.

[00117] The process flow **300** of FIG. **20** shall now be described to illustrate further aspects of FIG. **19**. Referring to block **302**, in order to determine the

IPR2025-01398
CrowdStrike EX1002 Page 183

causality of one or more key users with respect to cascades, social media data associated with cascades, in the form of the activity logs **206A-206C** or other such historical cascade event data, may be accessed by the computing device **202** from any one or more of the social media data sources **204A-204C**. The social media data may define a message associated with a cascade, which may include a video, a tweet, a news posting, a Facebook like, or any other form of social media that a user may adopt, broadcast/rebroadcast, or otherwise interact with. The data from the activity logs **206A-206C** may define tuples, which may be used as input parameters to be applied to causality metrics as described herein. In some embodiments, the activity logs **206A-206C** include a User ID, a message ID, and a timestamp.

[00118] Referring to block **304**, one or more key users may be identified for analysis. Key users may be defined as users of social media that interact with a cascade or adopt a social media message at an early stage; e.g., within a predetermined timeframe. Key users may be associated with social media user profiles that include at least a name, a possible photo, a URL, and a number of posts/tweets/retweets or other information indicating adoption behavior associated with the user profile, and location.

[00119] Referring to block **306**, with continuing reference to diagram **200**, a subset of the social media data accessed in block **302**, associated with the key users, may be applied to one or more causality metrics. Specifically, in some embodiments, the computing device **202** may be configured with at least one application **208**, which may define a plurality of modules as defined herein suitable for executing the functionality described herein, i.e. applying causality metrics **210** (as described in Table 3 above) to the data. Using the social media data, the application **208** generates a set of causality values for one or more key users. As shown in block **308**, a PSM account may be detected where the causality value of a given key user exceeds a predefined threshold. As further shown, and referenced in block **310**, the application **208** may include or otherwise be configured to implement time-decay extensions **212**, hyper-graph based machine learning (ML) **214**, and label propagation **216** to further define the data.

[00120] FIG. **21** is an example schematic diagram of a computing system **700** that may implement various methodologies discussed herein. For example, the computing system **700** may comprise a server used to execute aspects of the application **208**. The computing system **700** includes a bus **701** (i.e.,

IPR2025-01398
CrowdStrike EX1002 Page 184

interconnect), at least one processor **702** or other computing element, at least one communication port **703**, a main memory **704**, a removable storage media **705**, a read-only memory **706**, and a mass storage device **707**. Processor(s) **702** can be any known processor, such as, but not limited to, an Intel® Itanium® or Itanium 2® processor(s), AMD® Opteron® or Athlon MP® processor(s), or Motorola® lines of processors. Communication port **703** can be any of an RS-232 port for use with a modem based dial-up connection, a 10/100 Ethernet port, a Gigabit port using copper or fiber, or a USB port. Communication port(s) **703** may be chosen depending on a network such as a Local Area Network (LAN), a Wide Area Network (WAN), or any network to which the computer system **700** connects. Computing system may further include a transport and/or transit network **755**, a display screen **760**, an I/O port **740**, and an input device **745** such as a mouse or keyboard.

[00121] Main memory **704** can be Random Access Memory (RAM) or any other dynamic storage device(s) commonly known in the art. Read-only memory **706** can be any static storage device(s) such as Programmable Read-Only Memory (PROM) chips for storing static information such as instructions for processor **702**. Mass storage device **707** can be used to store information and instructions. For example, hard disks such as the Adaptec® family of Small Computer Serial Interface (SCSI) drives, an optical disc, an array of disks such as Redundant Array of Independent Disks (RAID), such as the Adaptec® family of RAID drives, or any other mass storage devices, may be used.

[00122] Bus **701** communicatively couples processor(s) **702** with the other memory, storage, and communications blocks. Bus **701** can be a PCI / PCI-X, SCSI, or Universal Serial Bus (USB) based system bus (or other) depending on the storage devices used. Removable storage media **705** can be any kind of external hard drives, thumb drives, Compact Disc – Read Only Memory (CD-ROM), Compact Disc – Re-Writable (CD-RW), Digital Video Disk – Read Only Memory (DVD-ROM), etc.

[00123] Embodiments herein may be provided as a computer program product, which may include a machine-readable medium having stored thereon instructions which may be used to program a computer (or other electronic devices) to perform a process. The machine-readable medium may include, but is not limited to optical discs, CD-ROMs, magneto-optical disks, ROMs, RAMs, erasable programmable read-only memories (EPROMs), electrically erasable programmable

IPR2025-01398
CrowdStrike EX1002 Page 185

read-only memories (EEPROMs), magnetic or optical cards, flash memory, or other type of media/machine-readable medium suitable for storing electronic instructions. Moreover, embodiments herein may also be downloaded as a computer program product, wherein the program may be transferred from a remote computer to a requesting computer by way of data signals embodied in a carrier wave or other propagation medium via a communication link (e.g., modem or network connection).

[00124] As shown, main memory **704** may be encoded with the application **208** that supports functionality discussed herein. In other words, aspects of the application **208** (and/or other resources as described herein) can be embodied as software code such as data and/or logic instructions (e.g., code stored in the memory or on another computer readable medium such as a disk) that supports processing functionality according to different embodiments described herein. During operation of one embodiment, processor(s) **702** accesses main memory **704** via the use of bus **701** in order to launch, run, execute, interpret or otherwise perform processes, such as through logic instructions, executing on the processor **702** and based on the application **208** stored in main memory or otherwise tangibly stored.

[00125] The description above includes example systems, methods, techniques, instruction sequences, and/or computer program products that embody techniques of the present disclosure. However, it is understood that the described disclosure may be practiced without these specific details. In the present disclosure, the methods disclosed may be implemented as sets of instructions or software readable by a device. Further, it is understood that the specific order or hierarchy of steps in the methods disclosed are instances of example approaches. Based upon design preferences, it is understood that the specific order or hierarchy of steps in the method can be rearranged while remaining within the disclosed subject matter. The accompanying method claims present elements of the various steps in a sample order, and are not necessarily meant to be limited to the specific order or hierarchy presented.

[00126] The described disclosure may be provided as a computer program product, or software, that may include a machine-readable medium having stored thereon instructions, which may be used to program a computer system (or other electronic devices) to perform a process according to the present disclosure. A machine-readable medium includes any mechanism for storing information in a form (e.g., software, processing application) readable by a machine (e.g., a computer).

IPR2025-01398
CrowdStrike EX1002 Page 186

The machine-readable medium may include, but is not limited to optical storage medium (e.g., CD-ROM); magneto-optical storage medium, read only memory (ROM); random access memory (RAM); erasable programmable memory (e.g., EPROM and EEPROM); flash memory; or other types of medium suitable for storing electronic instructions.

[00127] Certain embodiments are described herein as being implemented via one or more applications that include metrics which may collectively define modules. Such modules may be hardware-implemented, and may thus include at least one tangible unit capable of performing certain operations and may be configured or arranged in a certain manner. For example, a hardware-implemented module may comprise dedicated circuitry that is permanently configured (e.g., as a special-purpose processor, such as a field-programmable gate array (FPGA) or an application-specific integrated circuit (ASIC)) to perform certain operations. A hardware-implemented module may also comprise programmable circuitry (e.g., as encompassed within a general-purpose processor or other programmable processor) that is temporarily configured by software or firmware to perform certain operations. In some example embodiments, one or more computer systems (e.g., a standalone system, a client and/or server computer system, or a peer-to-peer computer system) or one or more processors may be configured by software (e.g., an application or application portion) as a hardware-implemented module that operates to perform certain operations as described herein.

[00128] Accordingly, the term "hardware-implemented module" or "module" encompasses a tangible entity, be that an entity that is physically constructed, permanently configured (e.g., hardwired), or temporarily configured (e.g., programmed) to operate in a certain manner and/or to perform certain operations described herein. Considering embodiments in which hardware-implemented modules are temporarily configured (e.g., programmed), each of the hardware-implemented modules need not be configured or instantiated at any one instance in time. For example, where the hardware-implemented modules comprise a general-purpose processor configured using software, the general-purpose processor may be configured as respective different hardware-implemented modules at different times. Software may accordingly configure a processor, for example, to constitute a particular hardware-implemented module at one instance of time and to constitute a different hardware-implemented module at a different instance of time.

IPR2025-01398
CrowdStrike EX1002 Page 187

[00129] Hardware-implemented modules may provide information to, and/or receive information from, other hardware-implemented modules. Accordingly, the described hardware-implemented modules may be regarded as being communicatively coupled. Where multiple of such hardware-implemented modules exist contemporaneously, communications may be achieved through signal transmission (e.g., over appropriate circuits and buses) that connect the hardware-implemented modules. In embodiments in which multiple hardware-implemented modules are configured or instantiated at different times, communications between such hardware-implemented modules may be achieved, for example, through the storage and retrieval of information in memory structures to which the multiple hardware-implemented modules have access. For example, one hardware-implemented module may perform an operation, and may store the output of that operation in a memory device to which it is communicatively coupled. A further hardware-implemented module may then, at a later time, access the memory device to retrieve and process the stored output. Hardware-implemented modules may also initiate communications with input or output devices.

[00130] It should be understood from the foregoing that, while particular embodiments have been illustrated and described, various modifications can be made thereto without departing from the spirit and scope of the invention as will be apparent to those skilled in the art. Such changes and modifications are within the scope and teachings of this invention as defined in the claims appended hereto.

IPR2025-01398
CrowdStrike EX1002 Page 188

CLAIMS

What is claimed is:

1. A method of detecting pathogenic social media accounts, comprising:

utilizing a processor in communication with a tangible storage medium storing instructions that are executed by the processor to perform operations comprising:

accessing data associated with a social media message adopted by a plurality of users during a predetermined time period, the data defining a cascade associated with the social media message;

identifying one or more key users from the plurality of users, the one or more key users having adopted the social media message before a predetermined number of the plurality of users; and

applying at least one causality metric to a subset of the data associated with the one or more key users to generate a set of causality values, wherein the set of causality values reflect causality associated with the cascade as exhibited by the one or more key users and further indicate whether the one or more key users are associated with a pathogenic social media account.

2. The method of claim 1, further comprising:

identifying a first user from the one or more key users, the first user having adopted the social media message before a predetermined number of the plurality of users; and

identifying a second user from the one or more key users, the second user having adopted the social media message subsequent to the first user and the second user being related to the first user.

3. The method of claim 2, wherein the cascade defines an action log of a plurality of tuples associated with the plurality of users where each of the

IPR2025-01398
CrowdStrike EX1002 Page 189

plurality of tuples represents when the social media message is adopted such that $(u, m, t) \in Actions$ corresponds with a respective user of the plurality of users designated u adopting the social media message designated $m \in M$ within the predetermined time period designated $t \in T$.

4. The method of claim 3, further comprising determining that the first user designated i precedes the second user designated j in the cascade where $t < t'$ and where $(i, m, t), (j, m, t') \in$ Action.

5. The method of claim 3, further comprising determining that the second user is related to the first user where:

$$R(i) = \{j \; s.t. \; j \neq i \; , \exists m \in M \; s.t. \; i, j \; are \; m-related\}$$

wherein R denotes relatedness.

6. The method of claim 2, further comprising:

(i) measuring a first probability that a size of the cascade meets a predetermined threshold given the first user adopts the social media message before the second user,

(ii) measuring a probability that the size of the cascade meets the predetermined threshold where the first user does not adopt the social media message before the second user, and

computing an average value of difference between (i) and (ii).

7. The method of claim 2, further comprising analyzing the cascade without influence by the first user to further assess causality for the cascade as exhibited by the first user.

8. The method of claim 1, wherein the at least one causality metric includes a Kleinberg-Mishra causality metric.

9. The method of claim 1, wherein the social media message was adopted when a user tweeted the social media message, retweeted the social media message, posted the social media message, liked the social media message, or viewed the social media message, and further wherein the social media

IPR2025-01398
CrowdStrike EX1002 Page 190

message includes one or more of a text, a picture, a video, audio, a URL or link to a website, or a hashtag.

10. The method of claim 1, further wherein the social media message is adopted by the plurality of users a predetermined number of times.

11. The method of claim 1, further comprising:

wherein the processor is configured to determine if a pathogenic social media account is active, wherein active indicates that the pathogenic social media account has not been suspended or deleted.

12. The method of claim 1, wherein the predetermined number of the plurality of users' actions is 10%, 20%, 30%, 40%, or 50%.

13. The method of claim 1, wherein the predetermined time period is at least 2 weeks.

14. The method of claim 1, further comprising representing each of the one or more key users by a causality vector $x \in \mathbb{R}^d$, computed by the at least one causality metric.

15. The method of claim 1, further comprising:

assigning different time points to the predetermined time period; and

assigning different weights to the set of causality values computed for the one or more key users, the different weights inversely proportional to distances from a given time *t* associated with the predetermined time period.

16. The method of claim 15, further comprising:

utilizing a sliding-time window to divide the predetermined time period into intervals;

deploying an exponential decay function of a form $f(x) = e^{-\alpha x}$ to account for a time-decay effect; and

IPR2025-01398
CrowdStrike EX1002 Page 191

computing an average between the set of causality values computed using the sliding-time window.

17. A system, comprising:

a computing device, configured to:

access data associated with a social media message, the data defining a cascade,

identify one or more key users from a plurality of users, the one or more key users having adopted the social media message before a predetermined number of the plurality of users; and

apply at least one causality metric to a subset of the data associated with the one or more key users to generate a set of causality values, wherein the set of causality values reflect causality associated with the cascade as exhibited by the one or more key users and further indicate whether the one or more key users are associated with a pathogenic social media account.

18. The system of claim 17, wherein the computing device is further configured to access a set of URLs associated with an action log of the plurality of users associated with cascades.

19. The system of claim 18, wherein the computing device is further configured to apply a form of multi-dimensional point process to the set of URLs to estimate an influence of nodes associated with the plurality of users.

20. The system of claim 19, wherein the multi-dimensional point process is a Hawkes process.

IPR2025-01398
CrowdStrike EX1002 Page 192

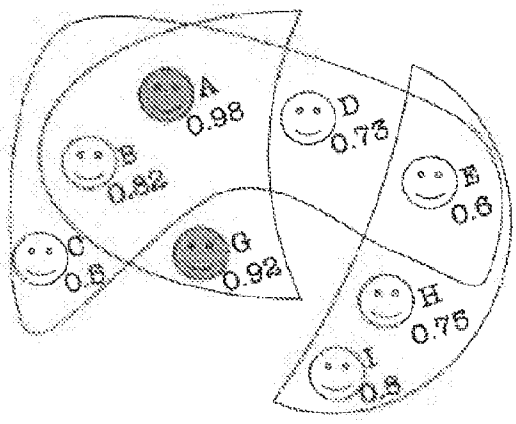


FIG. 1A

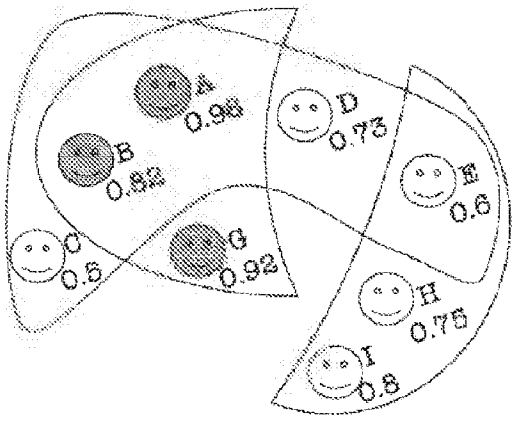


FIG. 1B

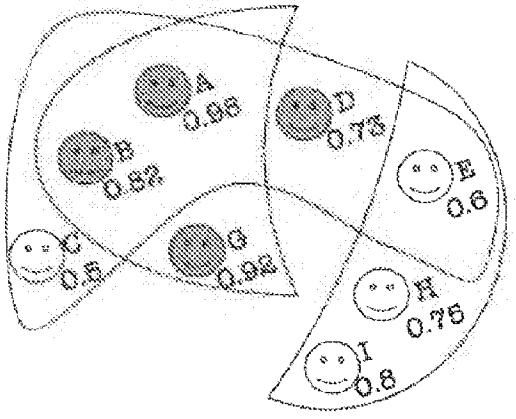


FIG. 1C

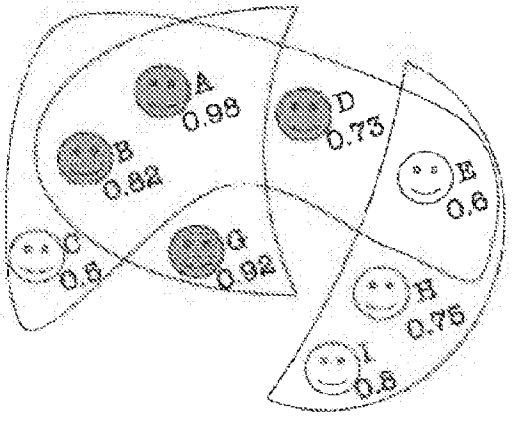


FIG. 1D

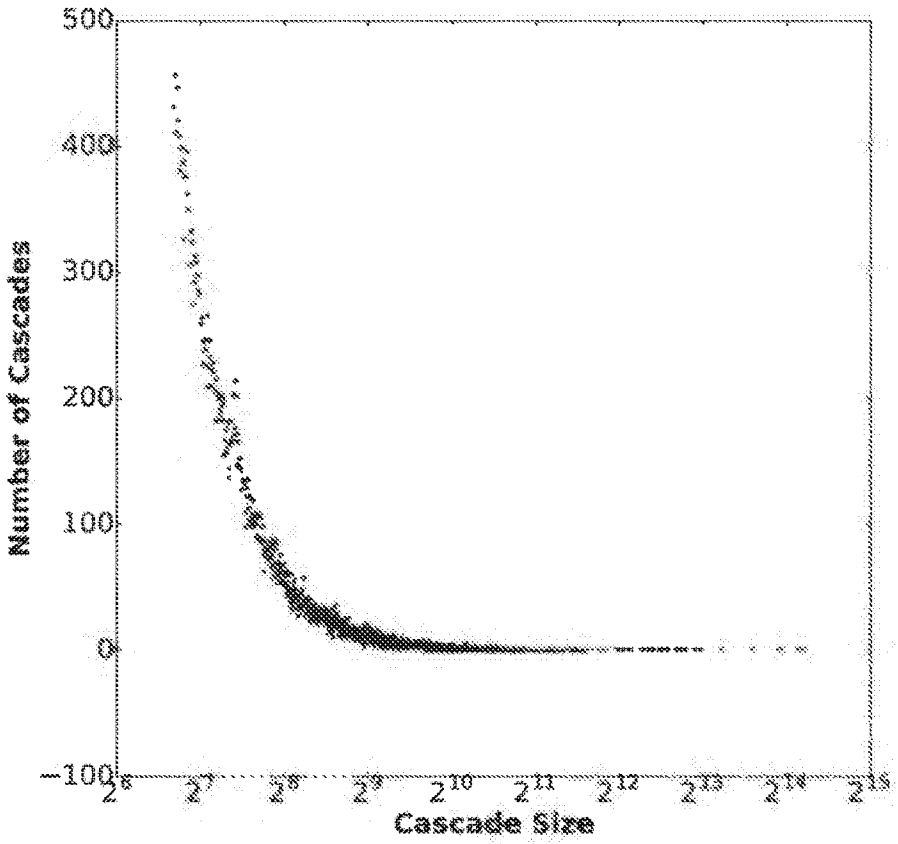


FIG. 2A

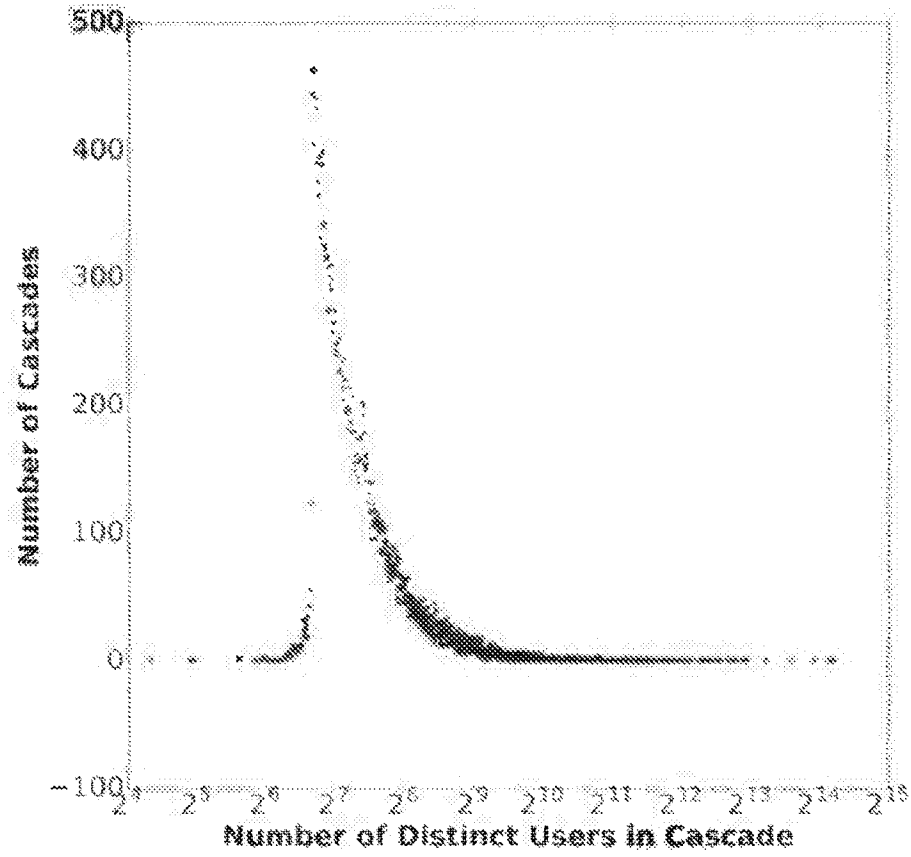


FIG. 2B

IPR2025-01398
CrowdStrike EX1002 Page 194

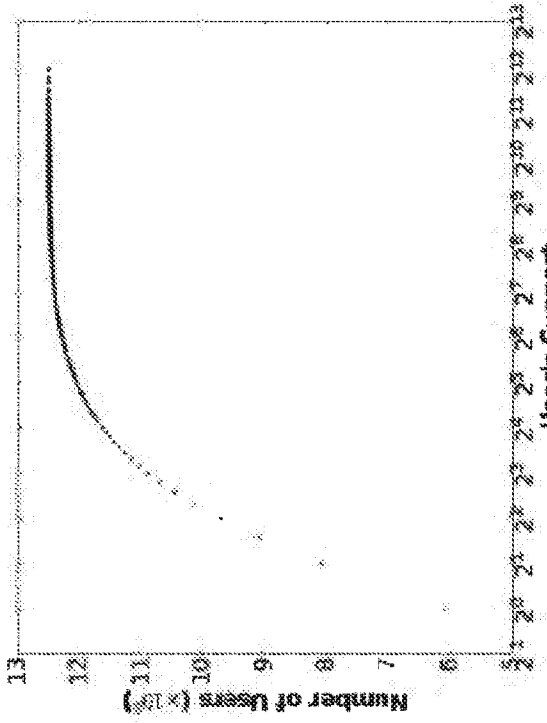


FIG. 3B

FIG. 3A

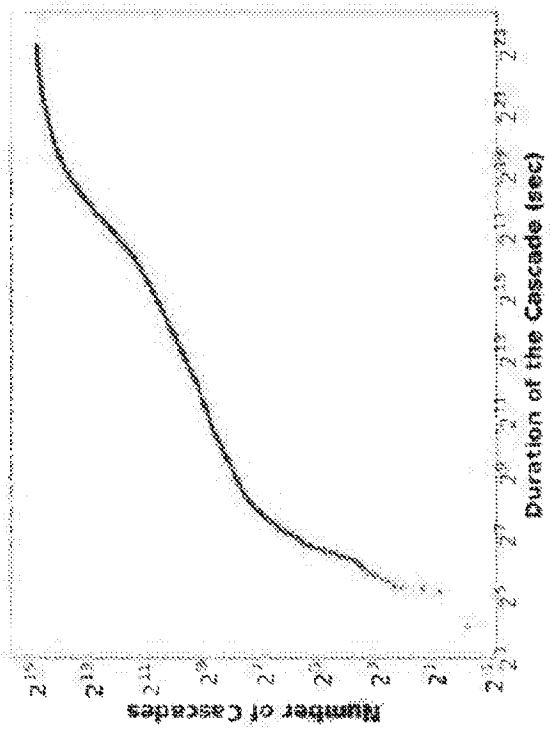
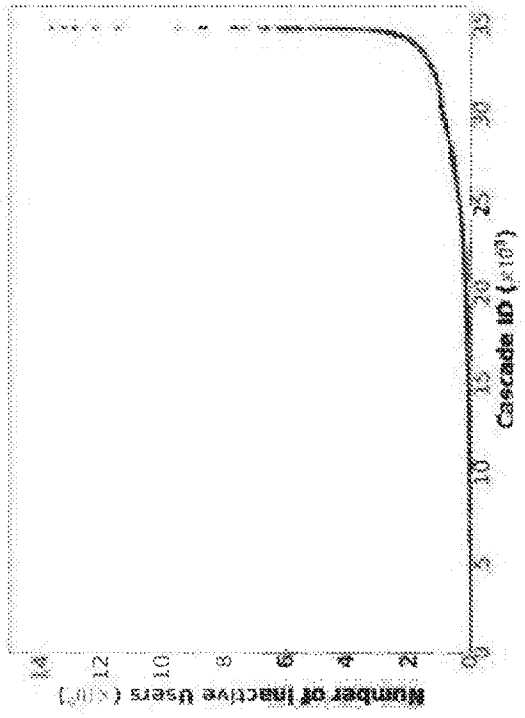
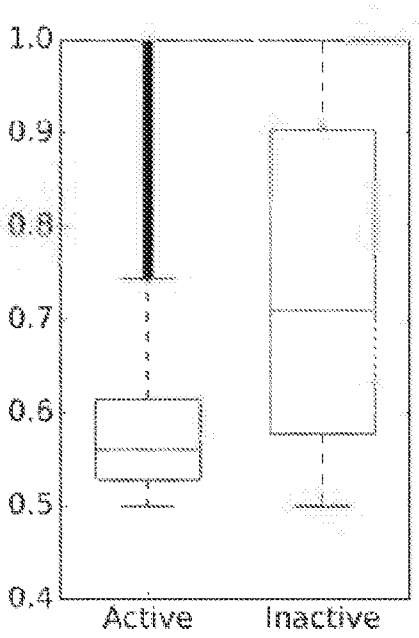


FIG. 3C

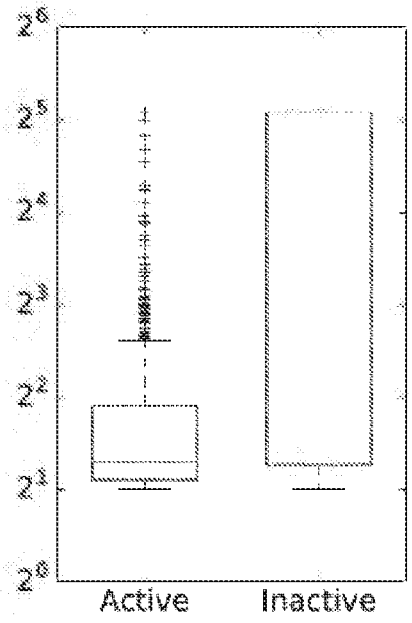


IPR2025-01398
CrowdStrike EX1002 Page 195



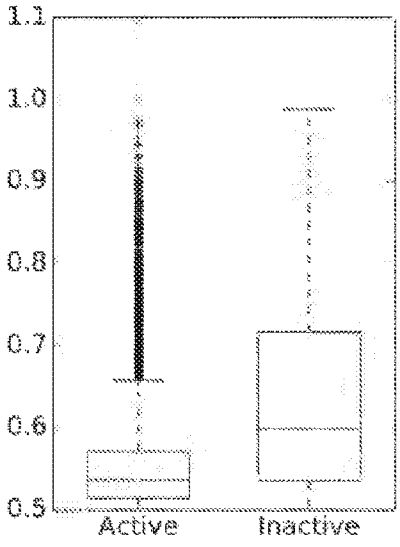
(a) $\epsilon_{K\&M} \geq 0.5$

FIG. 4A



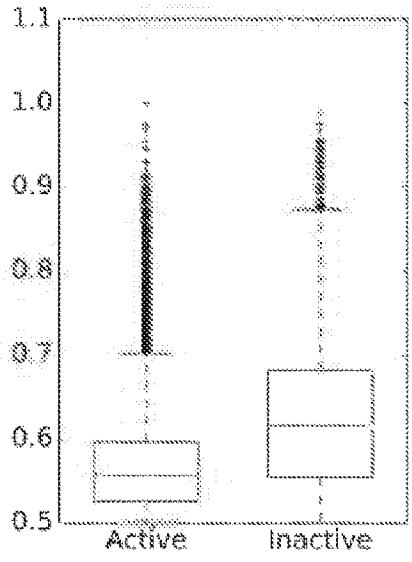
(b) $\epsilon_{rel} \geq 2$

FIG. 4B



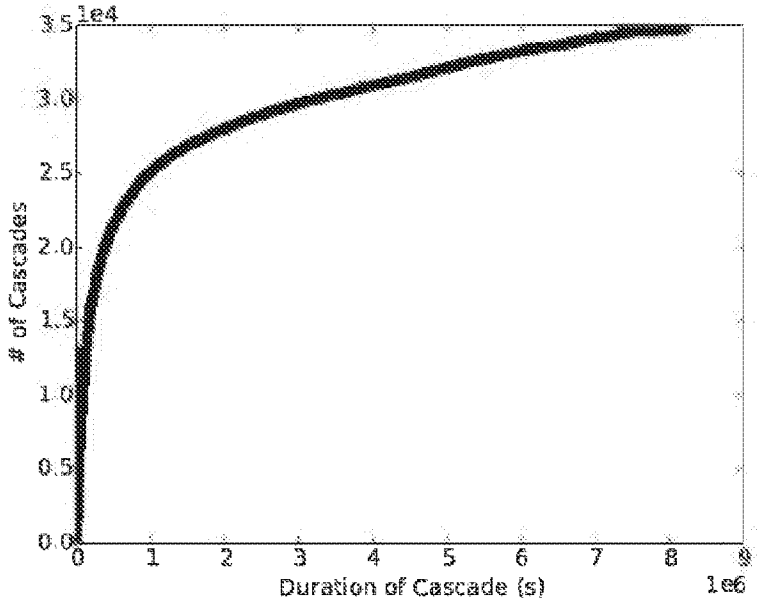
(c) $\epsilon_{nb} \geq 0.5$

FIG. 4C



(d) $\epsilon_{unb} \geq 0.5$

FIG. 4D


1e4
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
of Cascades
0
1
2
3
4
5
6
7
8
9
Duration of Cascade (s)
1e6

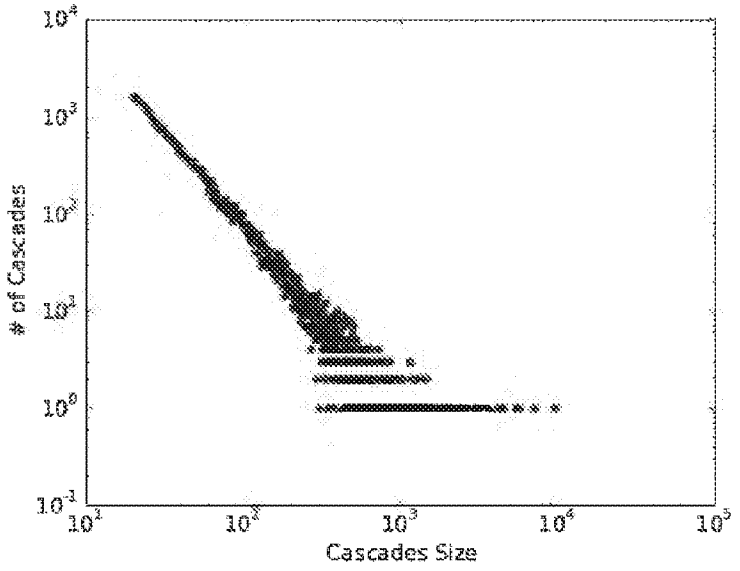

10^4
10^3
10^2
10^1
10^0
10^{-1}
of Cascades
10^1
10^2
10^3
10^4
10^5
Cascades Size

FIG. 5

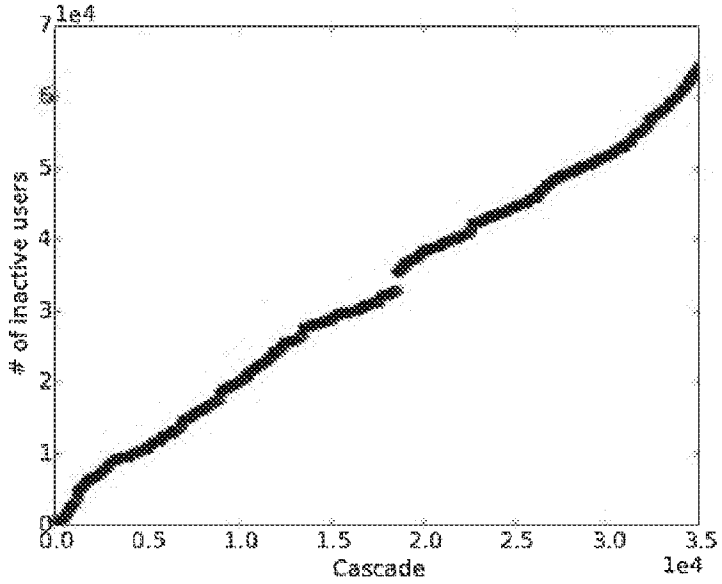

1e4
7
6
5
4
3
2
1
0
of inactive users
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Cascade
1e4

FIG. 6

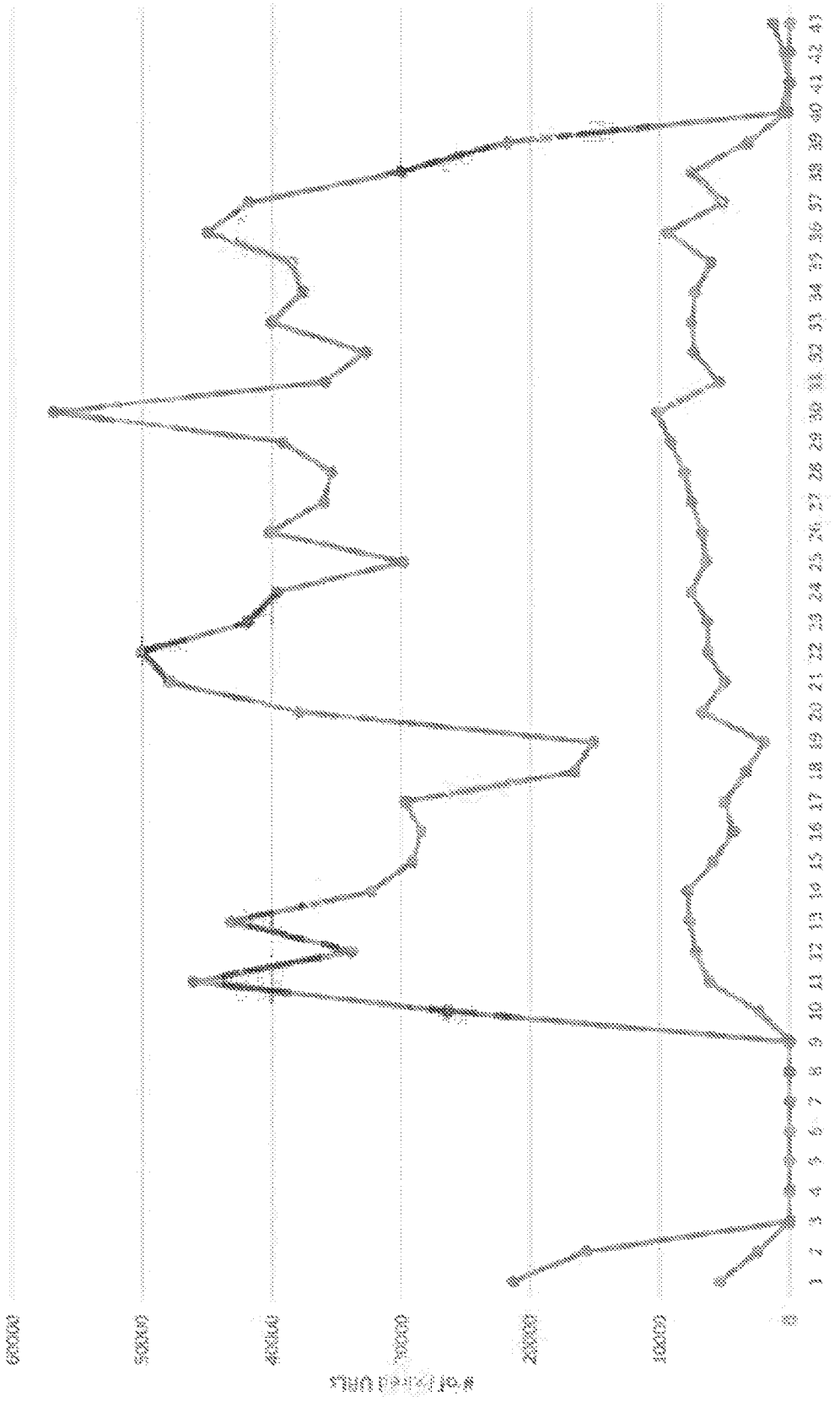


FIG. 7

IPR2025-01398
CrowdStrike EX1002 Page 199

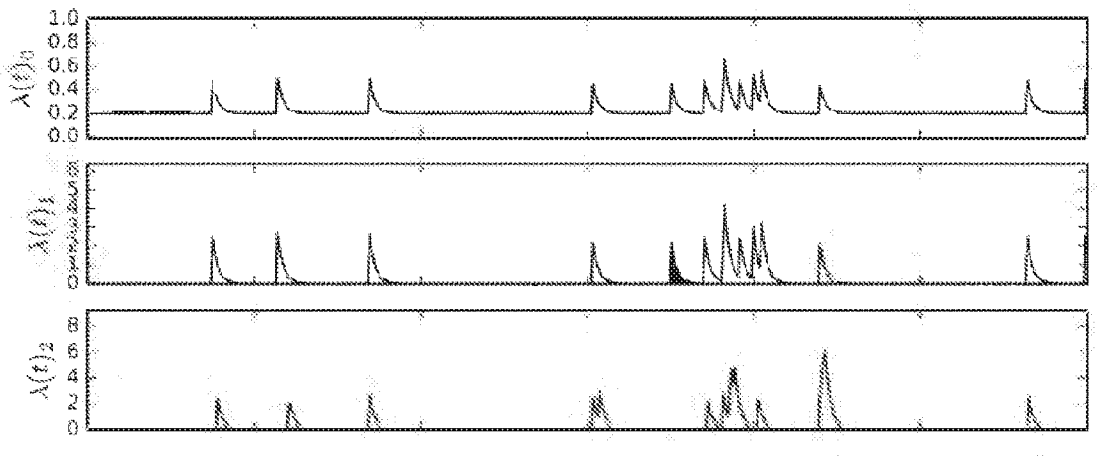


FIG. 8

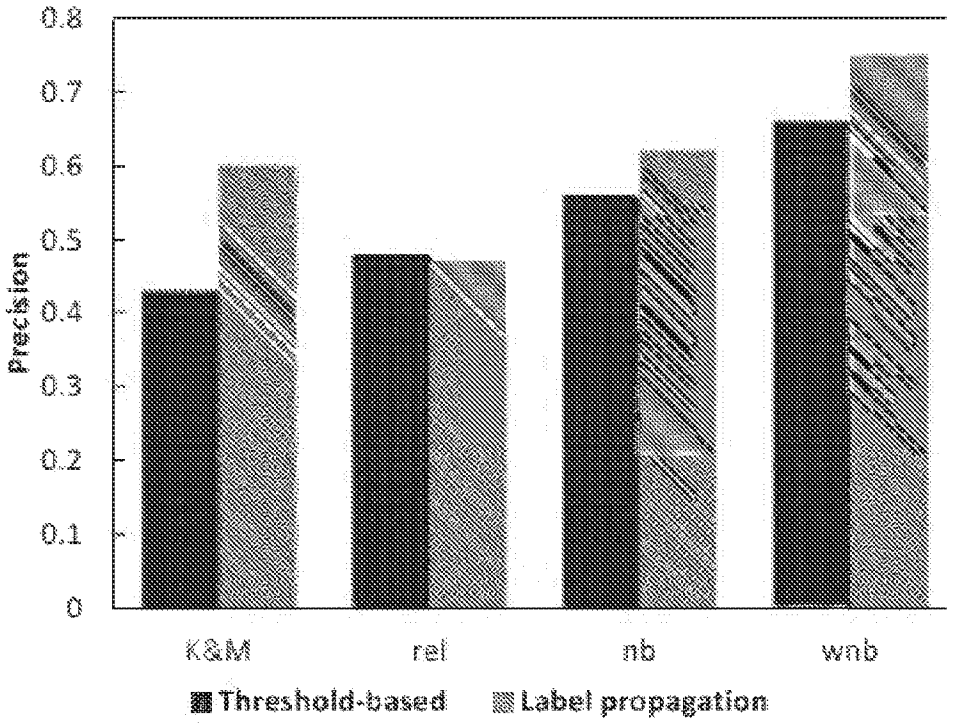

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Precision
K&M
rel
nb
wnb
Threshold-based
Label propagation

FIG. 9

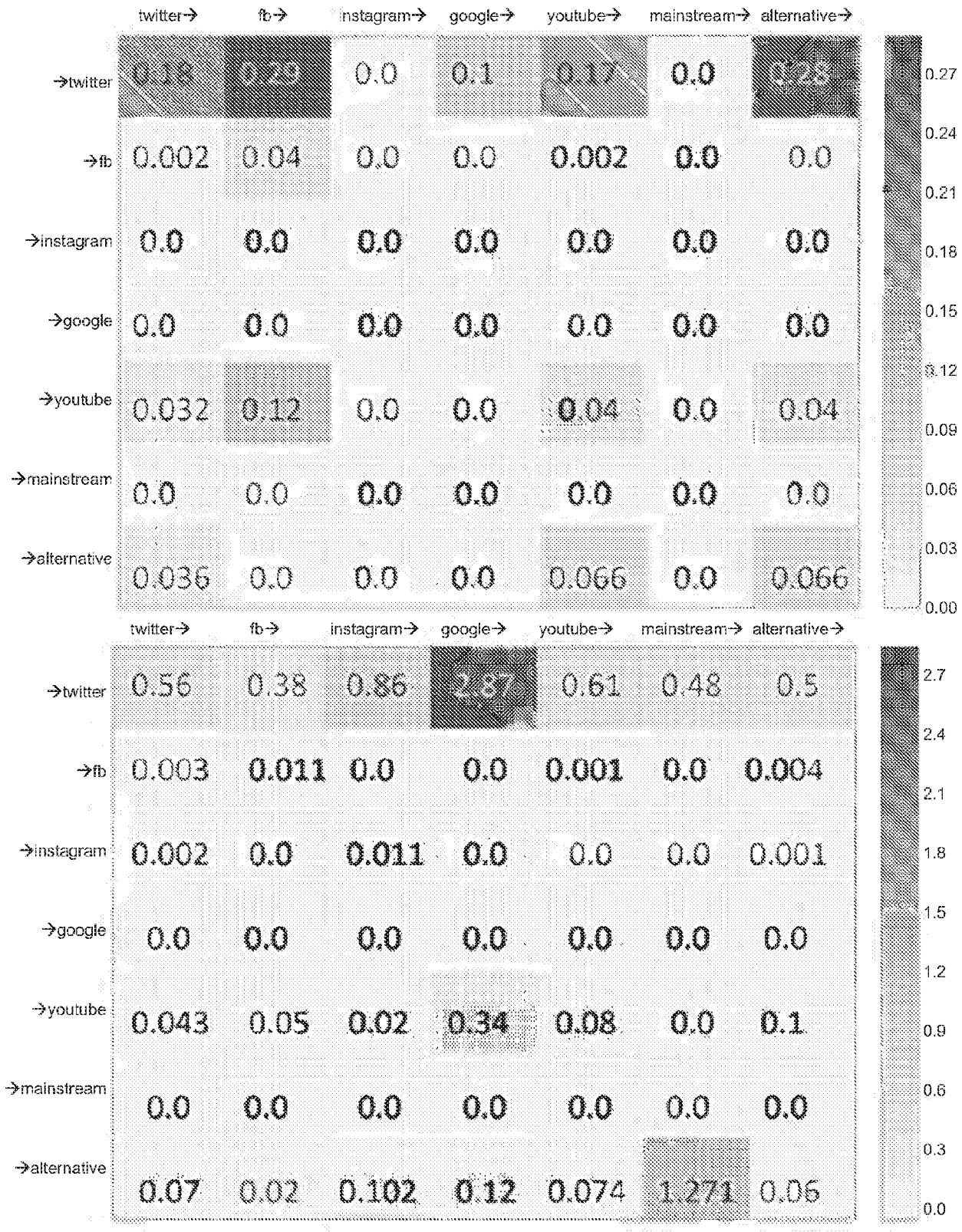


FIG. 10

IPR2025-01398
CrowdStrike EX1002 Page 202

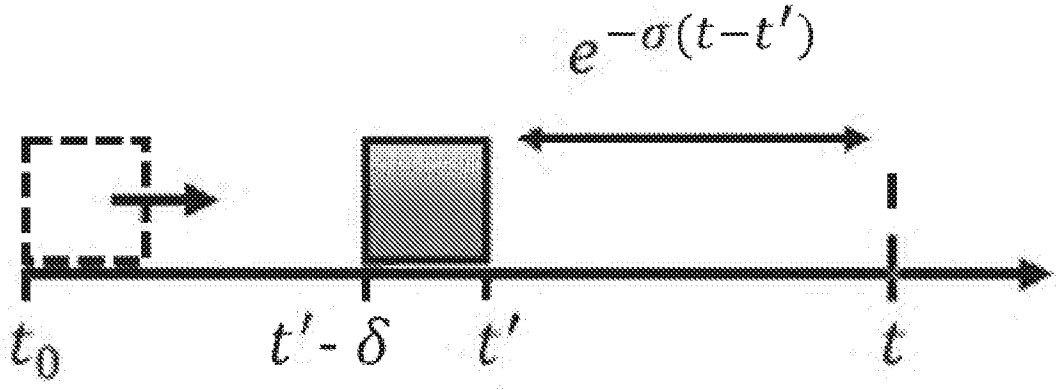


FIG. 11

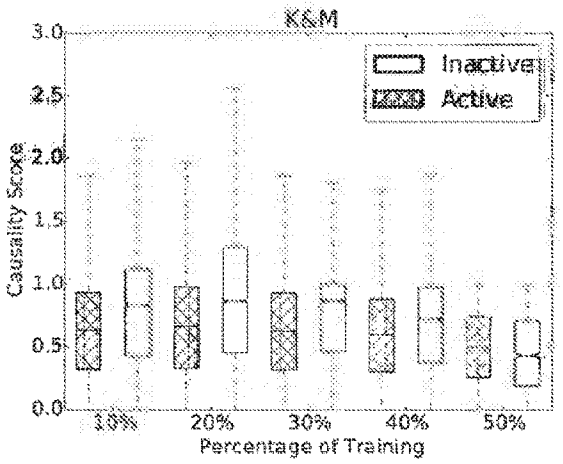


FIG. 12A

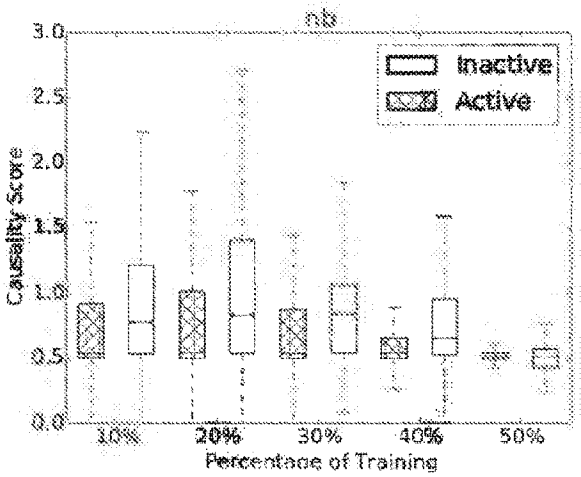


FIG. 12B

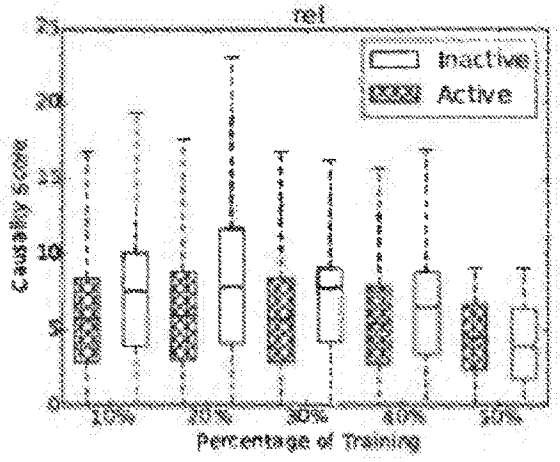


FIG. 12C

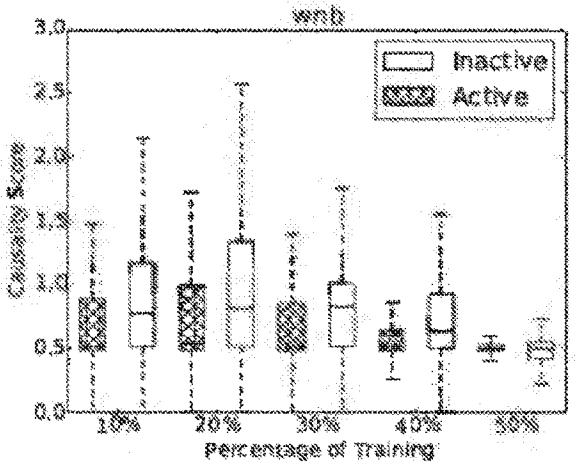


FIG. 12D

IPR2025-01398
CrowdStrike EX1002 Page 204

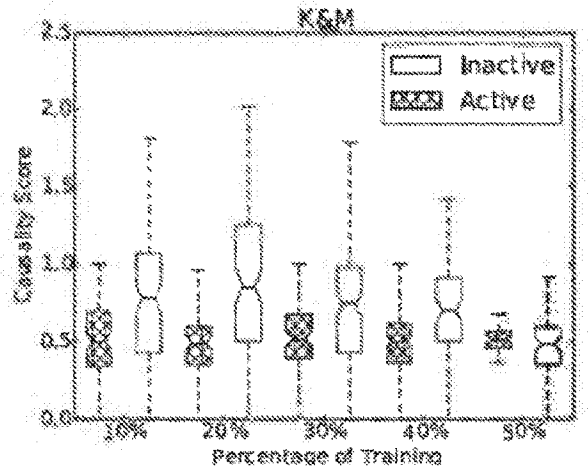


FIG. 13A

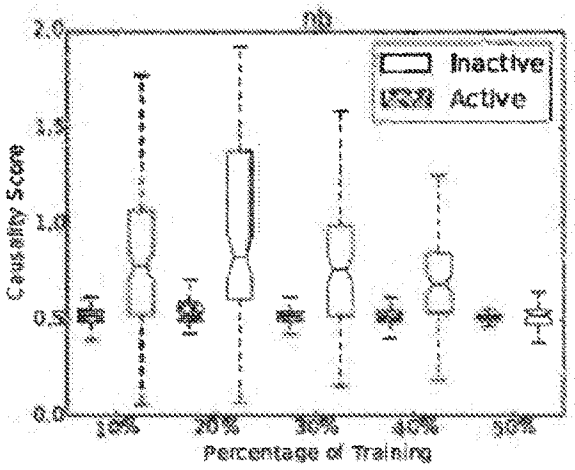


FIG. 13B

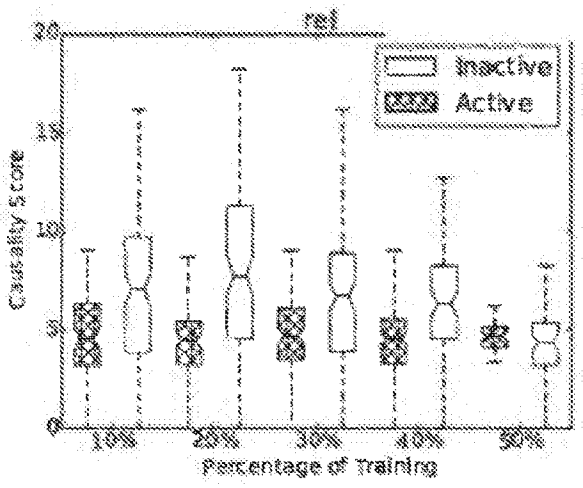


FIG. 13C

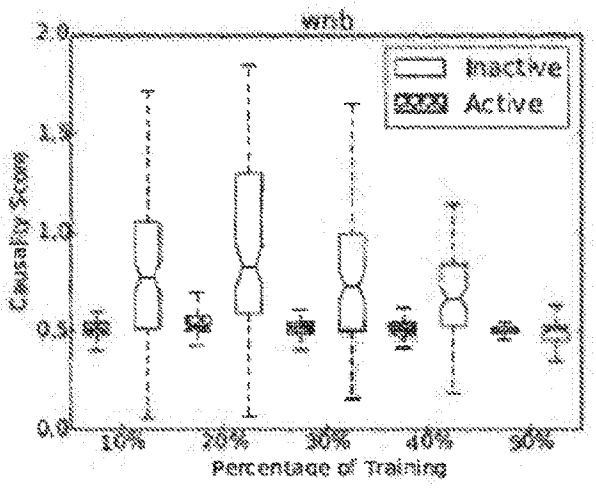


FIG. 13D

IPR2025-01398
CrowdStrike EX1002 Page 205

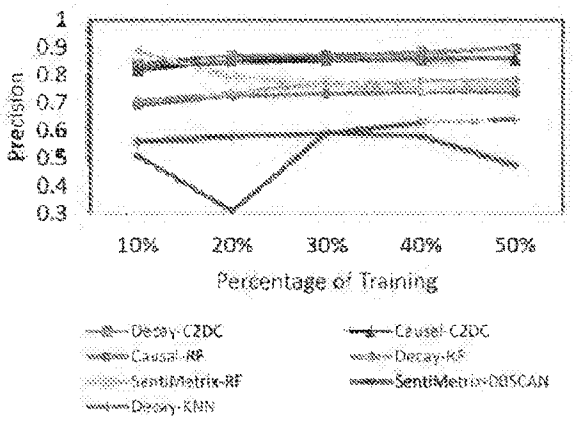


FIG. 14A

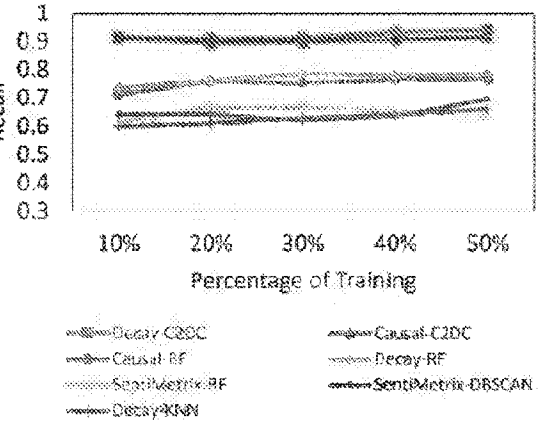


FIG. 14B

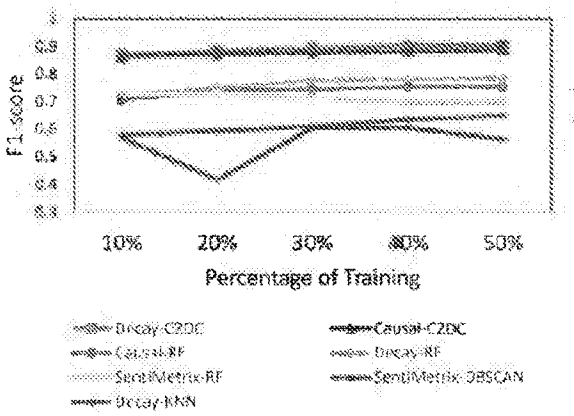


FIG. 15A

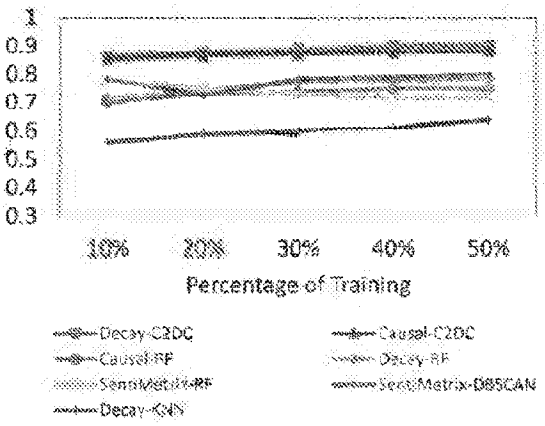


FIG. 15B

IPR2025-01398
CrowdStrike EX1002 Page 206

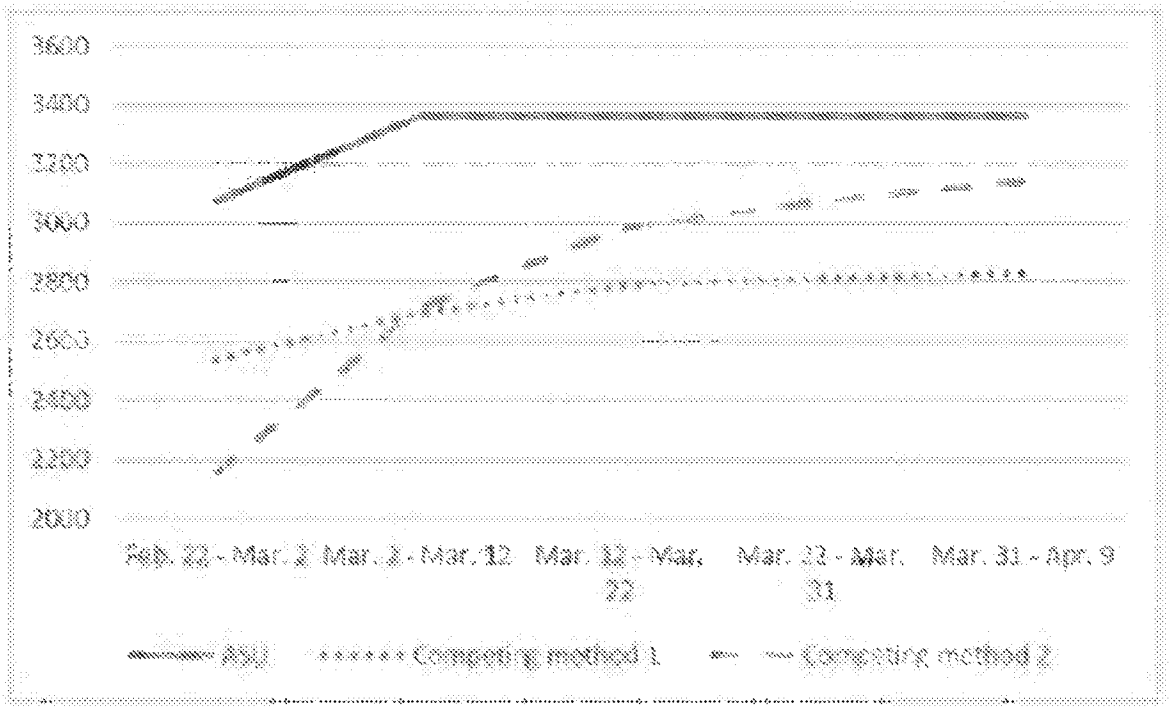


FIG. 16A

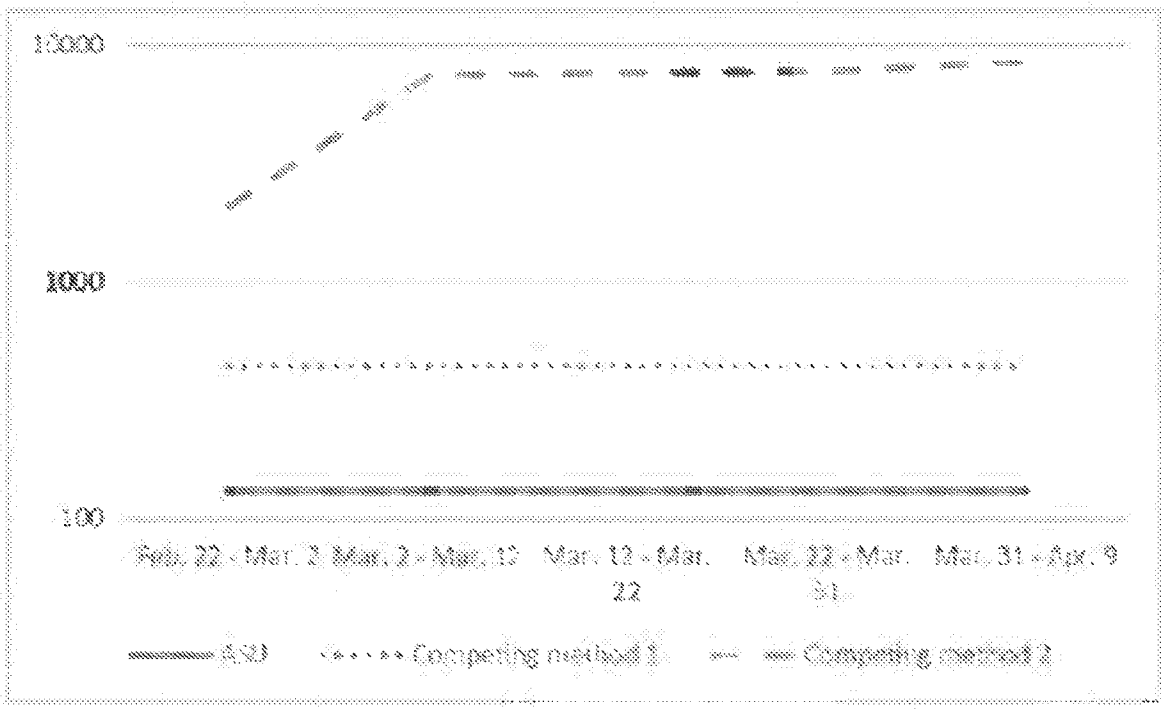


FIG. 16B

IPR2025-01398
CrowdStrike EX1002 Page 207

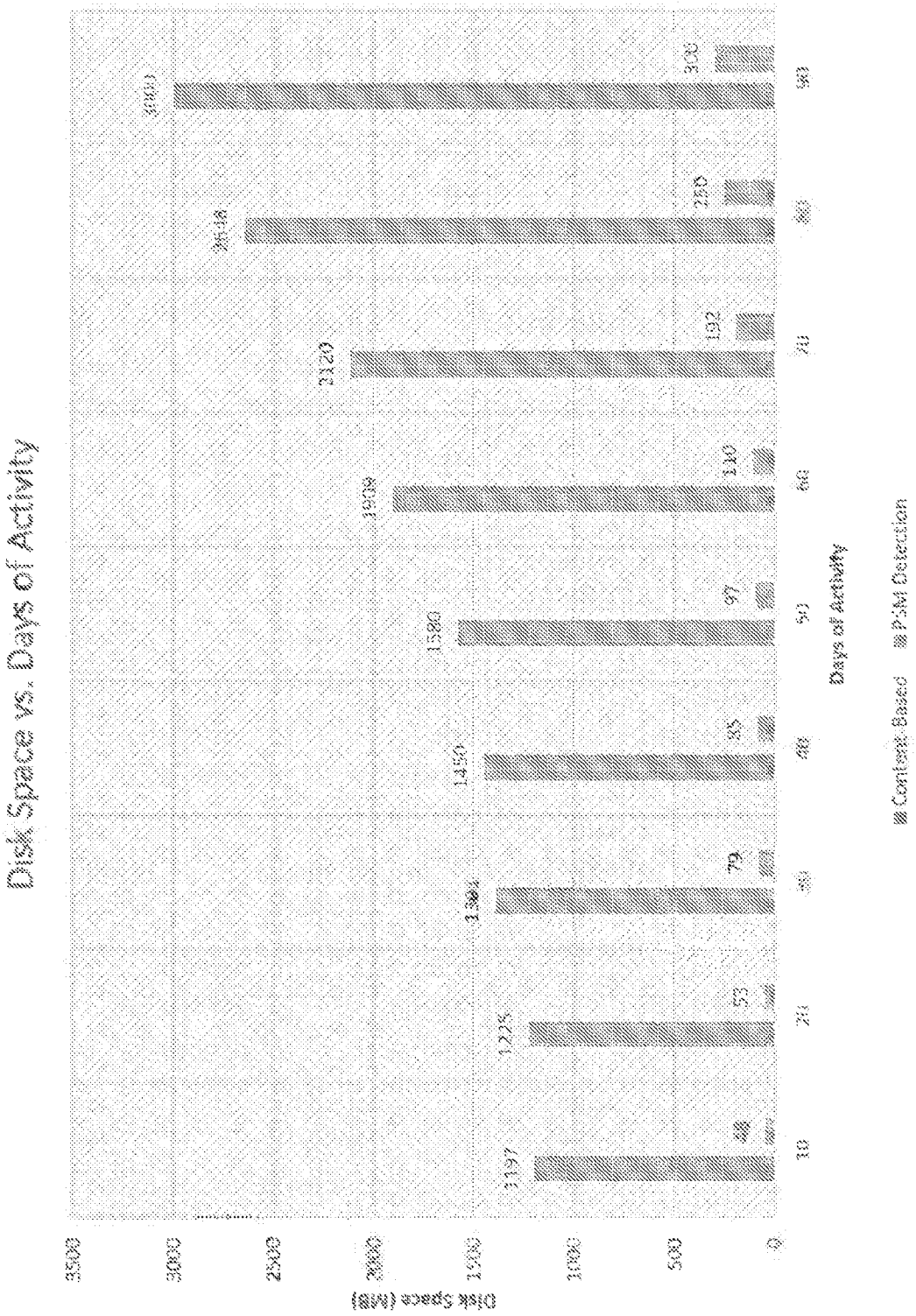

Disk Space vs. Days of Activity
Disk Space (MB)
3500
3000
2500
2000
1500
1000
500
0
3197
48
1225
53
1354
79
1450
85
1580
97
1909
140
3120
192
3548
255
3800
300
10
20
30
40
50
60
70
80
90
Days of Activity
Content-Based
PSM Detection

FIG. 17

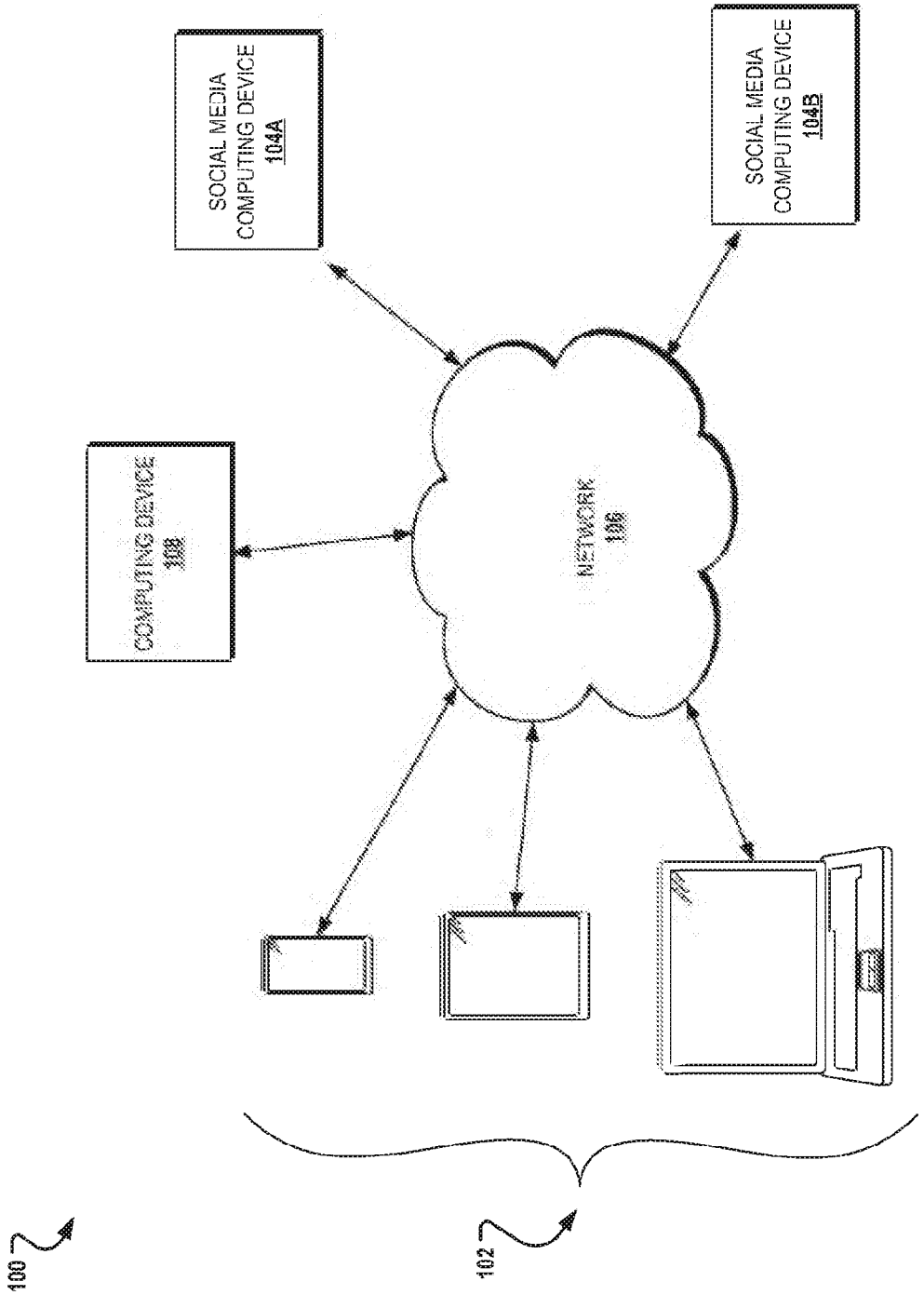

SOCIAL MEDIA COMPUTING DEVICE 104A
SOCIAL MEDIA COMPUTING DEVICE 104B
COMPUTING DEVICE 108
NETWORK 106
USER DEVICES
100
102

FIG. 18

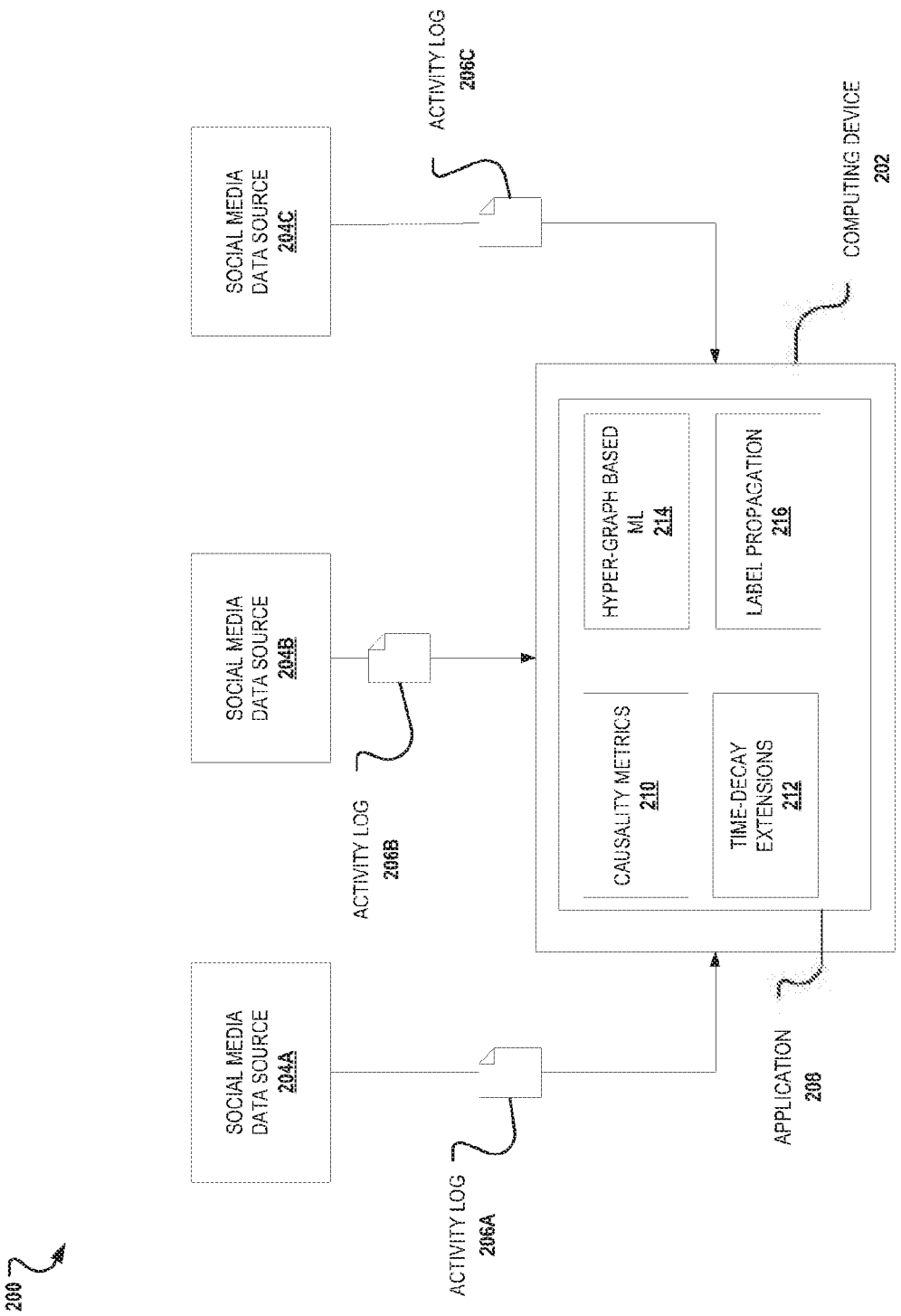

200
SOCIAL MEDIA DATA SOURCE 204A
SOCIAL MEDIA DATA SOURCE 204B
SOCIAL MEDIA DATA SOURCE 204C
ACTIVITY LOG 206A
ACTIVITY LOG 206B
ACTIVITY LOG 206C
COMPUTING DEVICE 202
HYPER-GRAPH BASED ML 214
LABEL PROPAGATION 216
CAUSALITY METRICS 210
TIME-DECAY EXTENSIONS 212
APPLICATION 208

FIG. 19

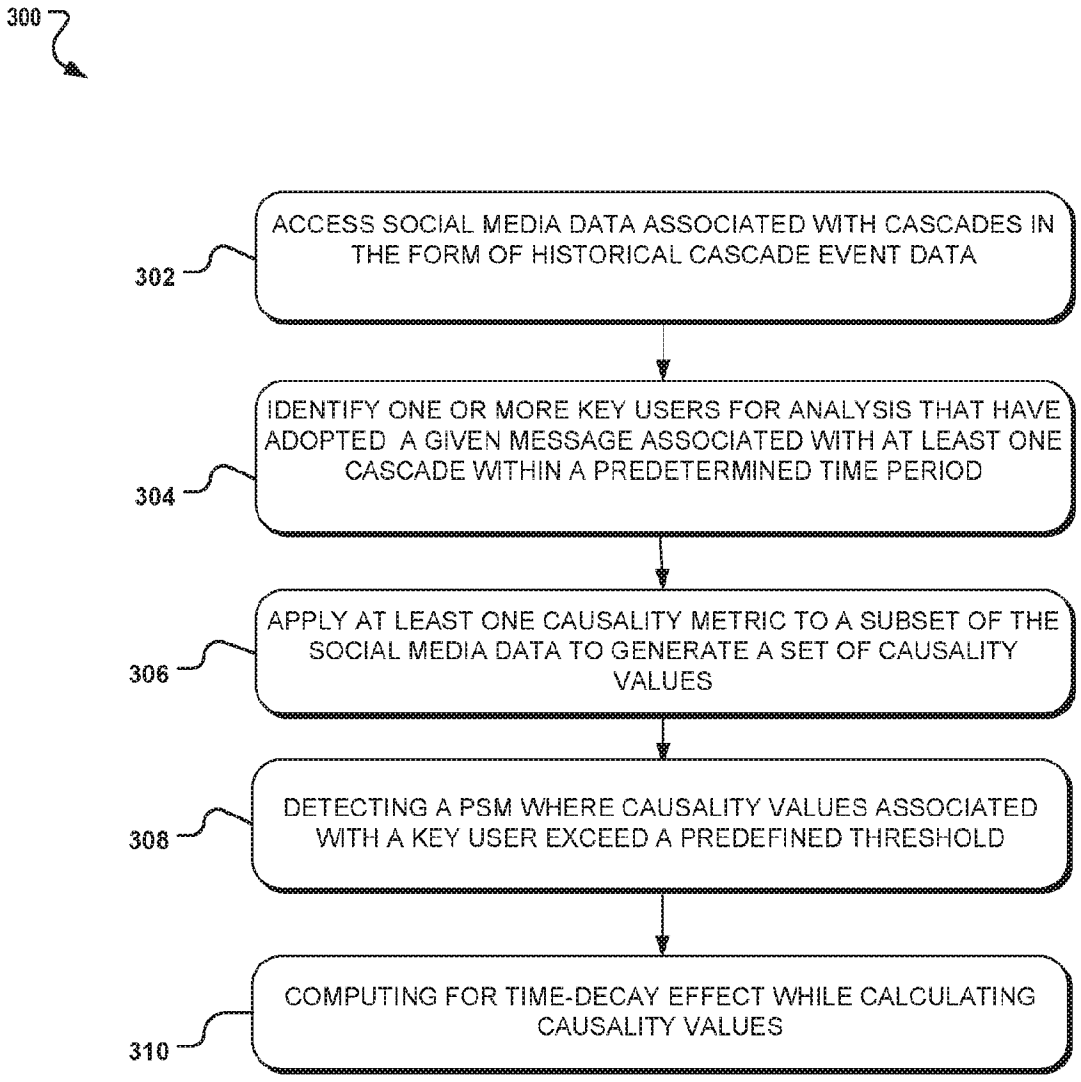

300
ACCESS SOCIAL MEDIA DATA ASSOCIATED WITH CASCADES IN THE FORM OF HISTORICAL CASCADE EVENT DATA
302
IDENTIFY ONE OR MORE KEY USERS FOR ANALYSIS THAT HAVE ADOPTED A GIVEN MESSAGE ASSOCIATED WITH AT LEAST ONE CASCADE WITHIN A PREDETERMINED TIME PERIOD
304
APPLY AT LEAST ONE CAUSALITY METRIC TO A SUBSET OF THE SOCIAL MEDIA DATA TO GENERATE A SET OF CAUSALITY VALUES
306
DETECTING A PSM WHERE CAUSALITY VALUES ASSOCIATED WITH A KEY USER EXCEED A PREDEFINED THRESHOLD
308
COMPUTING FOR TIME-DECAY EFFECT WHILE CALCULATING CAUSALITY VALUES
310

FIG. 20

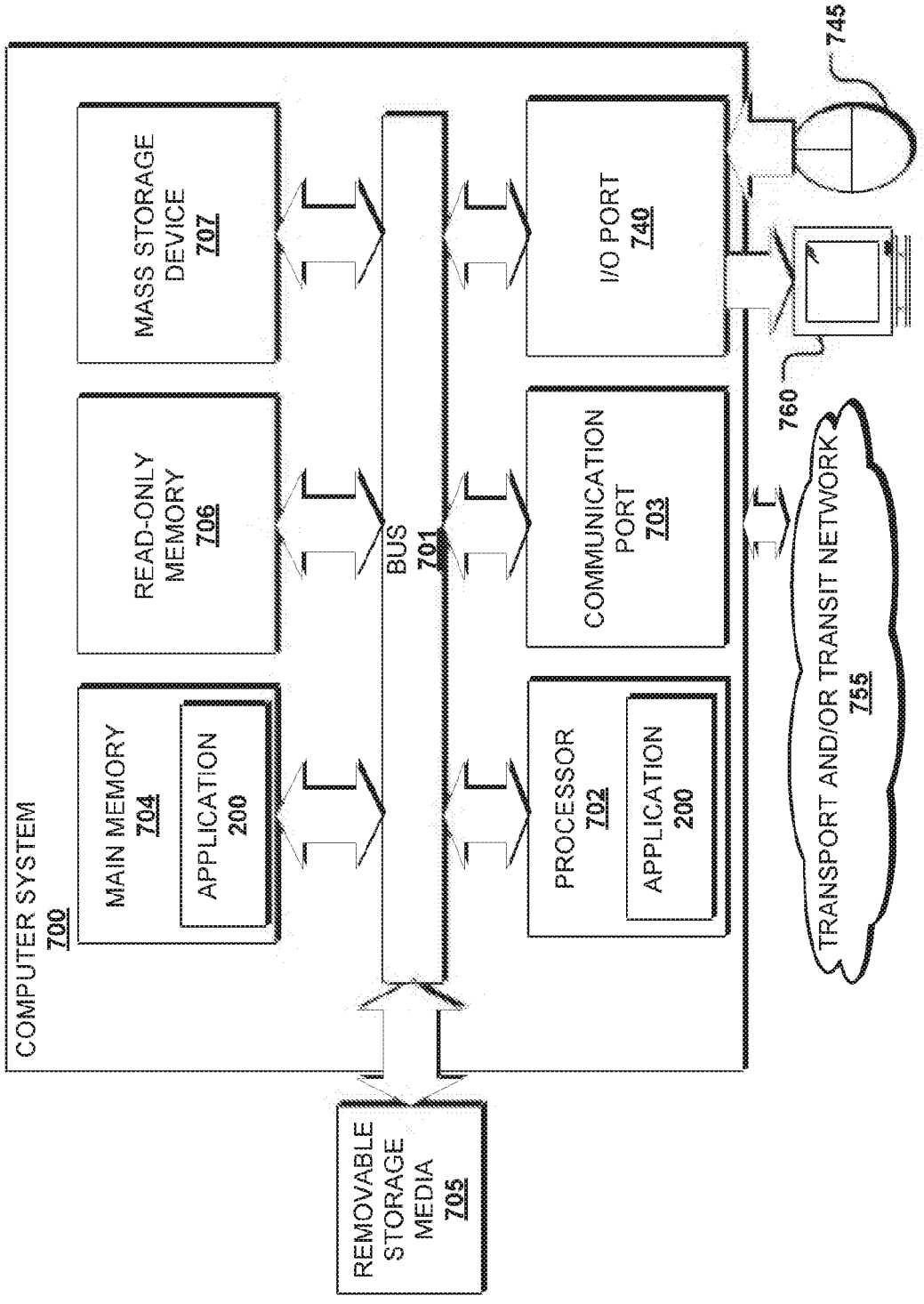

COMPUTER SYSTEM
700
MAIN MEMORY
704
APPLICATION
200
READ-ONLY
MEMORY
706
MASS STORAGE
DEVICE
707
BUS
701
PROCESSOR
702
APPLICATION
200
COMMUNICATION
PORT
703
I/O PORT
740
745
760
TRANSPORT AND/OR TRANSIT NETWORK
755
REMOVABLE
STORAGE
MEDIA
705

FIG. 21

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US19/17287

A. CLASSIFICATION OF SUBJECT MATTER
IPC - G06F 15/173, 16/24, 21/50 (2019.01)
CPC - G06F 15/173, 16/24, 21/50; G06Q 50/01; H04L 51/32

According to International Patent Classification (IPC) or to both national classification and IPC

B FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2014/0304343 A1 (AVAYA, INC.) 09 October 2014; entire document	1-20
A	US 8,955,127 B1 (SOKOLOV, I et al.) 10 February 2015; entire document	1-20
A	US 2014/0280610 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 18 September 2014; entire document	1-20
A	US 9,292,885 B2 (ASTORE, J) 22 March 2016; entire document	1-20
A	US 2017/0286671 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 05 October 2017; entire document	1-20

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
09 April 2019 (09.04.2019)	26 APR 2019

Name and mailing address of the ISA/	Authorized officer
Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Shane Thomas PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (January 2015)

Electronic Acknowledgement Receipt

EFS ID:	40295446
Application Number:	16405612
International Application Number:	
Confirmation Number:	4562
Title of Invention:	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB
First Named Inventor/Applicant Name:	Revanth Patil
Customer Number:	27148
Filer:	Ari M. Bai/Sam Bethune
Filer Authorized By:	Ari M. Bai
Attorney Docket Number:	055743-621457 (M18-191P)
Receipt Date:	17-AUG-2020
Filing Date:	07-MAY-2019
Time Stamp:	12:33:15
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Information Disclosure Statement (IDS) Form (SB08)	Information_Disclosure_Statement.pdf	98317 9411b64867ae0cd907eed532b7e2befc3e7398c7	no	6

Warnings:

Information:					
This is not an USPTO supplied IDS fillable form					
2	Non Patent Literature	ABBASI_2014.pdf	1975985 ea065bf70fe9aded805d4a500359ec4bd5c6b70a	no	8
Warnings:					
Information:					
3	Non Patent Literature	BREIMAN_2007.pdf	3406624 b48f6107aba31e813c815fcc719aabe13476a226	no	42
Warnings:					
Information:					
4	Non Patent Literature	CHAWLA_2002.pdf	3485216 fe95dd1dba4eccabf400eda54813385d0f8d4c40	no	37
Warnings:					
Information:					
5	Non Patent Literature	CHERTOFF_2015.pdf	1778541 af309b93af83f680169474d1ba14186a4657d7a2	no	18
Warnings:					
Information:					
6	Non Patent Literature	DEKEL_2010.pdf	292745 e502353147b2a93e16f52eda4f2b18d9263324fa	no	8
Warnings:					
Information:					
7	Non Patent Literature	HARRIS_1954.pdf	4349665 c260c45335ee59cbfb47e97c7ae27aeeef42105e	no	17
Warnings:					
Information:					
8	Non Patent Literature	LE_2014.pdf	343150 4081df889c3b35f6e70c07849034848d4b124854	no	9
Warnings:					
Information:					

9	Non Patent Literature	MANNING_2015.pdf	2358767 311ba16a4f384c49a302dcb15db6a3776855807d	no	10
Warnings:					
Information:					
10	Non Patent Literature	MARIN_2016.pdf	105003 66c943354a5ba8bafa02fef65fa78ef5b8ab4c42	no	3
Warnings:					
Information:					
11	Non Patent Literature	MOORE_2016.pdf	311176 c513f388b06afbbfd0bdb4c67528812823d3acc3	no	32
Warnings:					
Information:					
12	Non Patent Literature	NUNES_2016.pdf	421489 bb4473d242075c84db0063252bcc5d9b291e56ac	no	6
Warnings:					
Information:					
13	Non Patent Literature	PEDREGOSA_2011.pdf	182479 ddac74c766cf403f2f7fd28867101e87006f42af	no	6
Warnings:					
Information:					
14	Non Patent Literature	PEREZ-IGLESIAS_2009.pdf	253077 11f5993e3b3a71e23d4cd51c428c3228e527afdf	no	12
Warnings:					
Information:					
15	Non Patent Literature	RAJARAMAN_2011.pdf	280450 846baf7ed73a391c3c3eeb0ec387ee6e271ba5e5	no	32
Warnings:					
Information:					

16	Non Patent Literature	RONG_2014.pdf	287750 608bd60e701cd5f2fcb317619ce87976e2c473ac	no	19
Warnings:					
Information:					
17	Non Patent Literature	SAMTANI_2016.pdf	1674652 56320627f24ba6f2af1c1783e357a7fd1e6c4d2c	no	6
Warnings:					
Information:					
18	Non Patent Literature	SANDERSON_2009.pdf	161591 659ebd3b60080dbd1d14a7e41036c1e53ff57a3b	no	25
Warnings:					
Information:					
19	Non Patent Literature	SPARCK_1972.pdf	146562 8fae179bbbff2e1282d81eb8e736b8f96b0f5502	no	9
Warnings:					
Information:					
20	Non Patent Literature	TORPROJECT_2018.pdf	1068196 7358b97cb1e587b78d5682d7e71f780706b77f1a	no	10
Warnings:					
Information:					
21	Non Patent Literature	WIKIPEDIA_2017.pdf	1193020 dc780f6e6df544f04be03071bc82f2a823afe2d8	no	10
Warnings:					
Information:					
22	Non Patent Literature	WIKIPEDIA_2018.pdf	3520040 a5511fb376aa07e256b6d7a908b9bd385843150f	no	22
Warnings:					
Information:					

23	Foreign Reference	WO2019089389A1_.pdf	26109311 551b08851f7d7a81b8ede02154901e40ae2b9513	no	61
Warnings:					
Information:					
24	Foreign Reference	WO2019157335A1__.pdf	1616546 93923818b5a6a3f5881bf73f29cf417b164f1869	no	67
Warnings:					
Information:					
		Total Files Size (in bytes):	55420352		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/405,612	05/07/2019	Revanth Patil	055743-621457 (M18-191P)	4562

27148 7590 02/24/2021

POLSINELLI PC
900 WEST 48TH PLACE
SUITE 900
KANSAS CITY, MO 64112-1895

EXAMINER	
THOMAS-HOMESCU, ANNE L	

ART UNIT	PAPER NUMBER
2656	

NOTIFICATION DATE	DELIVERY MODE
02/24/2021	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

patentdocketing@polsinelli.com
rendsley@polsinelli.com

Office Action Summary	Application No. 16/405,612	Applicant(s) Patil et al.	
	Examiner ANNE L THOMAS-HOMESCU	Art Unit 2656	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTHS FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☑ Responsive to communication(s) filed on 07 May 2019.

☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2a)☐ This action is **FINAL.** 2b) ☑ This action is non-final.

3)☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

4)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims*

5) ☑ Claim(s) 1-20 is/are pending in the application.

5a) Of the above claim(s) _____ is/are withdrawn from consideration.

6) ☐ Claim(s) _____ is/are allowed.

7) ☑ Claim(s) 1-20 is/are rejected.

8) ☐ Claim(s) _____ is/are objected to.

9) ☐ Claim(s) _____ are subject to restriction and/or election requirement

* If any claims have been determined allowable, you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information, please see http://www.uspto.gov/patents/init_events/pph/index.jsp or send an inquiry to **PPHfeedback@uspto.gov.**

Application Papers

10)☐ The specification is objected to by the Examiner.

11)☑ The drawing(s) filed on 07 May 2019 is/are: a)☑ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

Certified copies:

a)☐ All b)☐ Some** c)☐ None of the:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

** See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☑ Notice of References Cited (PTO-892)

2) ☑ Information Disclosure Statement(s) (PTO/SB/08a and/or PTO/SB/08b) Paper No(s)/Mail Date _____.

3) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date _____.

4) ☐ Other: _____.

DETAILED ACTION

1. The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.

Information Disclosure Statement

1. The information disclosure statements (IDS) submitted on 17 August 2020 is in compliance with the provisions of 37 CFR 1.97. Accordingly, the information disclosure statement is being considered by the examiner.

Claim Rejections - 35 USC § 103

1. The following is a quotation of pre-AIA 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

1. **Claims 1-4 and 10-11** are rejected under pre-AIA 35 U.S.C. 103(a) as being obvious over "Hierarchical Classification of Web Content", hereinafter referred to as Dumais et al., in view of "Criminal motivation on the dark web: A categorization model for law enforcement", hereinafter referred to as Dalins et al., and further in view of US 20080189316, hereinafter referred to as Otsuka.

Regarding **claim 1**, Dumais et al. discloses a computer-implemented system for improving classification ~~of criminal activities~~ arising from cyber environments, comprising:

~~a processor ( ) configured to:~~

access a ground truth dataset generated from ~~deep~~ web forum information, the ground truth dataset defining a predetermined tag hierarchy including predetermined tag labels corresponding topics of the ~~deep~~ web forum information ("The basic idea behind our work is to use *classification* techniques to automatically organize search results into existing hierarchical structures. Classification models are learned offline using a training set of human-labeled documents and web categories," Dumais et al., p. 258, Web Data Set and Application section. Also, "During the training phase, a large set of web pages with known category labels are used to train a classifier," Dumais et al., p. 259, Text Classification Using SVMs section, 1st para. The labeled training documents are a ground truth dataset.);

access data related to a discussion topic for classification from a ~~deep~~ web forum ("During the testing or operational phase, the learned classifier is used to classify new web pages," Dumais et al., p. 259, Text Classification Using SVMs section, 1st para.);

extract a set of features from the data by assigning word vectors or paragraph vectors to the data (See Dumais et al., fig. 1. And, "After preprocessing, a binary vector was

created for each page indicating whether each term appeared in the page summary or not." Dumais et al., p.259, right col., 2nd para.); and

apply a machine classifier to the set of features to generate a prediction list of tags for classifying the discussion topic ("A support vector machine (SVM) algorithm was used as the classifier, because it has been shown in previous work to be both very fast and effective for text classification problems." Dumais et al., p. 259, Text Classification Using SVMs section, 2nd para.),

wherein the machine classifier is trained to maintain the predetermined tag hierarchy by regulating modification to parent and child tags associated with a predicted tag of the prediction list of tags based on a prediction probability of the predicted tag and predetermined tag thresholds assigned to the parent and child tags ("For the hierarchical models, p=0.2 for the multiplicative decision rule, and p_1=0.2 (first level) and p_2=0.5 (second level) for the Boolean decision rule." Dumais et al., p. 260, SVM Parameters section, 2nd para. Here, first level is the parent level/threshold corresponding to a first probability and second level is the child level/threshold corresponding to a second probability.).

Dumais et al., though, does not disclose classification of criminal activities arising from dark/deep web environments or a processor.

Dalins et al. is cited to disclose classification of criminal activities arising from dark/deep web environments ("In this paper we deploy the TMM by manually labelling a corpus of over 4000 unique Tor pages. We found a network impacted (but not dominated) by illicit

commerce and money laundering, bur almost completely devoid of violence and extremism. In short, criminality on this 'dark web' is based more upon greed and desire, rather than any particular political motivations," Dalins et al., Abstract. Also, "We labelled 2419 unique documents (by SHA-1 value) from the crawl, using the TMM. Due to the presence of excessively large (in terms of page count - refer Fig. 3) domains within Tor, we chose to report and measure the occurrences of categories on a *per* domain basis. Fig. 5 shows the relative occurrence of categories within Tor," p. 68, left col. See also Dalins et al., fig. 4 and 5.). *Dalins et al. benefits Dumais et al. by expanding the document classification techniques described in Dumais to classification of criminal activity on the dark web (Dalins et al., Abstract), thereby better tracking and anticipating criminal activities. Therefore, it would be obvious for one skilled in the art to combine the teachings of Dumais et al. with those of Dalins et al. to extend the application of the web classification method of Dumais to the law enforcement arena.*

Otsuka is cited to disclose a processor (Otsuka, para [0010], [0015], and [0023]. The examiner notes that a working computer, such as a workstation, PDS, etc. comprises a processor.). *Otsuka benefits Dumais et al. by providing a web crawler/spider for browsing the Web for suitable websites for classification (Otsuka, para [0023]). Therefore, it would be obvious for one skilled in the art to combine the teachings of Dumais et al. with those of Otsuka to more efficiently classify the web content of Dumais et al.*

Regarding **claim 2**, Dumais et al., as modified by Dalins et al. and Otsuka, discloses the computer-implemented system of claim 1, wherein the predetermined tag hierarchy is

divided into parent tags and child tags, wherein a topic represented by a child tag is representative of a subset of a topic represented by a parent tag ("In the hierarchical case, a model is learned to distinguish a second-level category from other categories within the same top level," Dumais et al., Abstract, 1st para. Also, "By utilizing known hierarchical structure, the classification problem can be decomposed into a set of smaller problems corresponding to hierarchical splits in the tree. Roughly speaking, one first learns to distinguish among classes at the top level, then lower level distinctions are learned only within the appropriate top level of the tree," Dumais et al., Introduction sec., 3rd para.), wherein each child tag corresponds to a particular parent tag, and each tag is assigned a corresponding probability value ("For the hierarchical models, p=0.2 for the multiplicative decision rule, and p_1=0.2 (first level) and p_2=0.5 (second level) for the Boolean decision rule," Dumais et al., p. 260, SVM Parameters section, 2nd para. Here, first level is the parent level/threshold corresponding to a first probability and second level is the child level/threshold corresponding to a second probability.).

Regarding **claim 3**, Dumais et al., as modified by Dalins et al. and Otsuka, discloses the computer-implemented system of claim 1, wherein the processor is configured to remove a child tag of the prediction list of tags corresponding to a parent tag from the prediction list of tags where a probability value associated with the child tag is below a predetermined threshold value ("...The former allows for matches that fall below individual thresholds, and the latter requires that the threshold be exceeded at every level. Both scoring rules allow items to be classified into multiple classes," Dumais et al., p. 261, left col., Second Level, Hierarchical sec.).

Regarding **claim 4**, Dumais et al., as modified by Dalins et al. and Otsuka, discloses the computer-implemented system of claim 1, wherein the processor is configured to add a parent tag corresponding to a child tag if a probability value associated with the parent tag is above a predetermined threshold value ("...The former allows for matches that fall below individual thresholds, and the latter requires that the threshold be exceeded at every level. Both scoring rules allow items to be classified into multiple classes," Dumais et al., p. 261, left col., Second Level, Hierarchical sec.).

Regarding **claim 10**, Dumais et al., as modified by Dalins et al. and Otsuka, discloses the computer-implemented system of claim 1, wherein the set of features are extracted from the data using a neural network, wherein the neural network is configured to learn a continuous distributed vector representation for documents and words within the data ("The efficiency of SVMs for both initial learning and real-time classification make them applicable to large dynamic collections like web content," Dumais et al., p. 258, left col., last para.).

Regarding **claim 11**, Dumais et al., as modified by Dalins et al. and Otsuka, discloses the computer-implemented system of claim 1, further comprising:

a crawler in operable communication with the processor and operable to traverse a plurality of ~~deep~~ web websites to retrieve Internet documents forming at least a portion of the ~~deep~~ web forum information ("Referring to FIG. 1, a noise word cleansing module 10 is configured to cleanse noise words from training data for a language.

Representative embodiments of this module include a text collector 11 configured to receive and/or retrieve text from any of a variety of sources… Additionally, or in the alternative, text collector 11 may include a web collection module 12 (shown in phantom) configured to collect web pages from websites of various categories. These web pages may be manually uploaded to text collector 11, or may be automatically uploaded, such as by use of one or more conventional web crawlers (also known as Web spiders or Web robots) that browse the Web for suitable websites," Otsuka, para [0023].}.

2. **Claim 5** is rejected under pre-AIA 35 U.S.C. 103(a) as being obvious over "Hierarchical Classification of Web Content", hereinafter referred to as Dumais et al., in view of "Criminal motivation on the dark web: A categorization model for law enforcement", hereinafter referred to as Dalins et al., further in view of US 20080189316, hereinafter referred to as Otsuka., further in view of US 20070250501, hereinafter referred to as Grubb et al.

Regarding **claim 5**, Dumais et al., as modified by Dalins et al. and Otsuka, discloses the computer-implemented system of claim 1, but not wherein the set of features are extracted using textual feature extraction to identify tags associated with topic titles of the data. Grubb et al. is cited to disclose wherein the set of features are extracted using textual feature extraction to identify tags associated with topic titles of the data ("All users, for each item that is saved, can specify metadata about the items including, but not limited to: title, tags, categories, topics, keywords, date, URL, referring URL, rating, comments, quotations from the item, author, publication, date, source, ISBN or ISSN,

library cataloging data, date stamps and/or bibliographic data," Grubb et al., [0032].).

Grubb et al. benefits Dumais et al. by tagging web content according to topic titles (Grubb et al., [0032]), thereby allowing a user to quickly ascertain a distribution of relevant topics. Therefore, it would be obvious to one skilled in the art to combine the teachings of Dumais et al. with those of Grubb et al. to improve the web content classification of Dumais et al.

3. **Claim 6** is rejected under pre-AIA 35 U.S.C. 103(a) as being obvious over "Hierarchical Classification of Web Content", hereinafter referred to as Dumais et al., in view of "Criminal motivation on the dark web: A categorization model for law enforcement", hereinafter referred to as Dalins et al., further in view of US 20080189316, hereinafter referred to as Otsuka., further in view of US 20070250501, hereinafter referred to as Grubb et al.

Regarding **claim 6**, Dumais et al., as modified by Dalins et al. and Otsuka, discloses the computer-implemented system of claim 1, but not wherein a set of synthetic data samples are created for addition to the ground truth dataset to address class imbalance by creating a set of synthetic sample points along a line defined between a first minority feature vector and a second feature vector and adding the set of synthetic sample points to the ground truth dataset. Chawla et al. is cited to disclose wherein a set of synthetic data samples are created for addition to the ground truth dataset to address class imbalance by creating a set of synthetic sample points along a line defined between a first minority feature vector and a second feature vector and adding the set of synthetic sample points to the ground truth dataset ("We generate synthetic examples in

a less application-specific manner by operating "feature space" rather than "data space". The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/ all of the *k* minority class nearest neighbors," Chawla et al., sec. 4.2, 1st para. See also Chawla et al., fig. 3.). *Chawla et al. benefits Dumais et al. by performing a biased sampling of clusters to obtain a diverse balanced sample (Chawla et al., sec. 4.2, 1st para). Therefore, it would be obvious for one skilled in the art to combine the teachings of Dumais et al. with those of Chawla et al. to improve the classification scheme of Dumais et al.*

4. **Claim 7** is rejected under pre-AIA 35 U.S.C. 103(a) as being obvious "Hierarchical Classification of Web Content", hereinafter referred to as Dumais et al., in view of "Criminal motivation on the dark web: A categorization model for law enforcement", hereinafter referred to as Dalins et al., further in view of US 20080189316, hereinafter referred to as Otsuka., further in view of A Task Level Metric for Measuring Web Search Satisfaction and its Application on Improving Relevance Estimation, hereinafter referred to as Hassan et al.

Regarding **claim 7**, Dumais et al., as modified by Dalins et al. and Otsuka, discloses the computer-implemented system of claim 1, but not wherein the ground truth dataset is supplemented by, for every sample of the ground truth dataset in a minority class, adding top similar documents from the ~~deep~~ web forum information. Hassan et al. is cited to disclose wherein the ground truth dataset is supplemented by, for every sample of the ground truth dataset in a minority class, adding top similar documents from the ~~deep~~ web forum information ("More than 83% of our data was for goals where users

ended up being satisfied with their search experiment. Hence the class distribution was not balanced. Imbalance in the class distribution often causes machine learning algorithms to perform poorly on the minority class [25]. In our case, we are more interested in the minority (DSAT) class. Hence, the cost of misclassifying the minority class is usually higher than the cost of misclassifying the majority class. Several researchers have experimentally evaluated the benefit of using sampling to artificially balance imbalanced data [4, 18, 26]. A common way of addressing class imbalance is to artificially rebalance the training data," Hassan et al., p. 131, left col., 2nd para.), as identified using an elastic search similarity score, and adding them to the ground truth dataset (Hassan et al., fig. 5 shows use of BM25, an elastic similarity metric, as described in the applicant's specification, para [0050].). *Hassan et al. benefits Dumais et al. by providing a metric to measure the relevance of search results (Hassan et al., sec. 7.2). Therefore, it would be obvious for one skilled in the art to combine the teachings of Dumais et al. with those of Hassan et al. to improve the web content classification method of Dumais et al.*

5. **Claims 8 and 9** are rejected under pre-AIA 35 U.S.C. 103(a) as being obvious over "Hierarchical Classification of Web Content", hereinafter referred to as Dumais et al., in view of "Criminal motivation on the dark web: A categorization model for law enforcement", hereinafter referred to as Dalins et al., further in view of US 20080189316, hereinafter referred to as Otsuka., further in view of US 20170193060, hereinafter referred to as Khanzode et al.

Regarding **claim 8**, Dumais et al., as modified by Dalins et al. and Otsuka, discloses the computer-implemented system of claim 1, but not wherein the data is parsed and segmented into tokens, and predetermined characters and predetermined common words are removed from the data. Khanzode et al. is cited to disclose wherein the data is parsed and segmented into tokens, and predetermined characters and predetermined common words are removed from the data ("In some embodiments, an analyzer may also include one or more filters that transform a search query before, during, and/or after tokenization. For example, an analyzer may transform a string entirely into lower-case, perform stemming and/or lemmatization (e.g., replacing "ran" or "running" with "run" and/or using synonyms), remove common words (e.g., "the," "is," and/or "of"), and/or remove markup (e.g., removing hypertext markup language characters), Khanzode et al., para [0034].). *Khanzode et al. benefits Dumais et al. by analyzing search queries with a tokenizer that separates out tokens (i.e., strings) from a search query and correlates those tokens with the appropriate fields or portions of fields (Khanzode et al., para [0034]). Therefore, it would be obvious for one skilled in the art to combine the teachings of Dumais et al. with those of Khanzode et al. to improve the web content search of Dumais et al.*

Regarding **claim 9**, Dumais et al., as modified by Dalins et al. and Otsuka, discloses the computer-implemented system of claim 1, but not wherein words within the data are reduced to their root form using stemming and lemmatization procedures. Khanzode et al. is cited to disclose wherein words within the data are reduced to their root form using stemming and lemmatization procedures ("In some embodiments, an analyzer may also

include one or more filters that transform a search query before, during, and/or after tokenization. For example, an analyzer may transform a string entirely into lower-case, perform stemming and/or lemmatization (e.g., replacing "ran" or "running" with "run" and/or using synonyms), remove common words (e.g., “the,” “is,” and/or “of”), and/or remove markup (e.g., removing hypertext markup language characters), Khanzode et al., para [0034].). *Khanzode et al. benefits Dumais et al. by analyzing search queries with a tokenizer that separates out tokens (i.e., strings) from a search query and correlates those tokens with the appropriate fields or portions of fields (Khanzode et al., para [0034]). Therefore, it would be obvious for one skilled in the art to combine the teachings of Dumais et al. with those of Khanzode et al. to improve the web content search of Dumais et al.*

6. **Claims 12, 14, and 15** are rejected under pre-AIA 35 U.S.C. 103(a) as being obvious over “Hierarchical Classification of Web Content”, hereinafter referred to as Dumais et al., in view of “SMOTE: Synthetic Minority Over-sampling Technique”, hereinafter referred to as Chawla et al., and further in view of US 20180246983, hereinafter referred to as Rathod.

Regarding **claim 12**, Dumais et al. discloses a method, comprising:

accessing a plurality of documents from ~~deep~~ web forums, each of the plurality of documents defining a topic (“During the testing or operational phase, the learned classifier is used to classify new web pages,” Dumais et al., p. 259, Text Classification Using SVMs section, 1st para.);

generating a ground truth dataset from the data, the ground truth dataset defining a predetermined tag hierarchy including predetermined tag labels corresponding to the topic of each of the plurality of documents ("The basic idea behind our work is to use *classification* techniques to automatically organize search results into existing hierarchical structures. Classification models are learned offline using a training set of human-labeled documents and web categories," Dumais et al., p. 258, Web Data Set and Application section. Also, "During the training phase, a large set of web pages with known category labels are used to train a classifier," Dumais et al., p. 259, Text Classification Using SVMs section, 1st para. The labeled training documents are a ground truth dataset.);

~~modifying the ground truth dataset in view of a minority class to address class imbalance; and~~

generating a machine classifier to classify additional documents from the ~~deep~~ web forums ("A support vector machine (SVM) algorithm was used as the classifier, because it has been shown in previous work to be both very fast and effective for text classification problems," Dumais et al., p. 259, Text Classification Using SVMs section, 2nd para.), the machine classifier configured to maintain the predetermined tag hierarchy defined by the ground truth dataset ("For the hierarchical models, p=0.2 for the multiplicative decision rule, and p_1=0.2 (first level) and p_2=0.5 (second level) for the Boolean decision rule," Dumais et al., p. 260, SVM Parameters section, 2nd para. Here, first level is the parent level/threshold corresponding to a first probability and second level is the child level/threshold corresponding to a second probability.).

Dumais et al., though, does not disclose modifying the ground truth dataset in view of a minority class to address class imbalance a *deep* web forum, or a processor for performing the above steps.

Chawla et al. is cited to disclose modifying the ground truth dataset in view of a minority class to address class imbalance ("An approach to t he construction of classifiers from imbalanced datasets is described," Chawla et al., Abstract. Also, "We generate synthetic examples in a less application-specific manner by operating "feature space" rather than "data space". The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/ all of the *k* minority class nearest neighbors," Chawla et al., sec. 4.2, 1st para. See also Chawla et al., fig. 3.). *Chawla et al. benefits Dumais et al. by performing a biased sampling of clusters to obtain a diverse balanced sample (Chawla et al., sec. 4.2, 1st para). Therefore, it would be obvious for one skilled in the art to combine the teachings of Dumais et al. with those of Chawla et al. to improve the classification scheme of Dumais et al.*

Neither Dumais et al. nor Chawla et al., though, disclose a minority class to address class imbalance a *deep* web forum, or a processor for performing the above steps.

Rathod is cited to disclose a *deep* web forum ("Internet content that is not capable of being searched by a web search engine is generally described as the deep web. Web search engine," Rathod, para [0326].) and a processor for performing the above steps (Rathod, para [0206]). *Rathod benefits Dumais et al. by expanding the document*

classification techniques described in Dumais to deep web forums (Rathod, para [0326]). Therefore, it would be obvious for one skilled in the art to combine the teachings of Dumais et al. with those of Rathod to extend the application of the hierarchical classification of Dumais to deep web content.

Regarding **claim 14**, Dumais et al., as modified by Chawla et al. and Rathod, discloses the method of claim 12, wherein a line is defined between a minority class feature vector and a nearest neighbor feature vector ("We generate synthetic examples in a less application-specific manner by operating "feature space" rather than "data space". The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/ all of the *k* minority class nearest neighbors," Chawla et al., sec. 4.2, 1st para. See also Chawla et al., fig. 3.}.

Regarding **claim 15**, Dumais et al., as modified by Chawla et al. and Rathod, discloses the method of claim 14, wherein a synthetic minority sample is added to the ground truth dataset by selecting a random point along the line defined between the minority class feature vector and the nearest neighbor feature vector is chosen and adding the random point to the minority class feature vector ("We generate synthetic examples in a less application-specific manner by operating "feature space" rather than "data space". The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/ all of the *k* minority class nearest neighbors…This causes the selection of a random point along the line

segment between two specific features. This approach effectively forces the decision region of the minority class to become more general," Chawla et al., sec. 4.2, 1st para. See also Chawla et al., fig. 3.).

7. **Claim 13** is rejected under pre-AIA 35 U.S.C. 103(a) as being obvious over "Hierarchical Classification of Web Content", hereinafter referred to as Dumais et al., in view of "SMOTE: Synthetic Minority Over-sampling Technique", hereinafter referred to as Chawla et al., further in view of US 20180246983, hereinafter referred to as Rathod, and further in view of "A Task Level Metric for Measuring Web Search Satisfaction and its Application on Improving Relevance Estimation", hereinafter referred to as Hassan et al.

Regarding **claim 13**, Dumais et al., as modified by Chawla et al. and Rathod, discloses the method of claim 12, but not further comprising modifying the ground truth dataset by introducing an additional document topically similar to the minority class, the additional document found by assigning an elastic search similarity score to each of the plurality of documents. Hassan et al. is cited to disclose modifying the ground truth dataset by introducing an additional document topically similar to the minority class ("More than 83% of our data was for goals where users ended up being satisfied with their search experiment. Hence the class distribution was not balanced. Imbalance in the class distribution often causes machine learning algorithms to perform poorly on the minority class [25]. In our case, we are more interested in the minority (DSAT) class. Hence, the cost of misclassifying the minority class is usually higher than the cost of misclassifying the majority class. Several researchers have experimentally evaluated the benefit of

using sampling to artificially balance imbalanced data [4, 18, 26]. A common way of addressing class imbalance is to artificially rebalance the training data," Hassan et al., p. 131, left col., 2nd para.), the additional document found by assigning an elastic search similarity score to each of the plurality of documents (Hassan et al., fig. 5 shows use of BM25, an elastic similarity metric, as described in the applicant's specification, para [0050].). *Hassan et al. benefits Dumais et al. by providing a metric to measure the relevance of search results (Hassan et al., sec. 7.2). Therefore, it would be obvious for one skilled in the art to combine the teachings of Dumais et al. with those of Hassan et al. to improve the web content classification method of Dumais et al.*

8. **Claim 16** is rejected under pre-AIA 35 U.S.C. 103(a) as being obvious over "Hierarchical Classification of Web Content", hereinafter referred to as Dumais et al., in view of "SMOTE: Synthetic Minority Over-sampling Technique", hereinafter referred to as Chawla et al., further in view of US 20180246983, hereinafter referred to as Rathod, and further in view of US 20080189316, hereinafter referred to as Otsuka.

Regarding **claim 16**, Dumais et al., as modified by Chawla et al. and Rathod, discloses the method of claim 12, but not further comprising generating the ground truth dataset by leveraging a spider program to crawl one or more ~~deep~~ web forums and retrieve HTML pages. Otsuka is cited to disclose generating the ground truth dataset by leveraging a spider program to crawl one or more ~~deep~~ web forums and retrieve HTML pages ("Referring to FIG. 1, a noise word cleansing module 10 is configured to cleanse noise words from training data for a language. Representative embodiments of this module include a text collector 11 configured to receive and/or retrieve text from any of

a variety of sources…Additionally, or in the alternative, text collector 11 may include a web collection module 12 (shown in phantom) configured to collect web pages from websites of various categories. These web pages may be manually uploaded to text collector 11, or may be automatically uploaded, such as by use of one or more conventional web crawlers (also known as Web spiders or Web robots) that browse the Web for suitable websites," Otsuka, para [0023].). *Otsuka benefits Dumais et al. by providing a web crawler/spider for browsing the Web for suitable websites for classification (Otsuka, para [0023]). Therefore, it would be obvious for one skilled in the art to combine the teachings of Dumais et al. with those of Otsuka to more efficiently classify the web content of Dumais et al.*

9. **Claim 16** is rejected under pre-AIA 35 U.S.C. 103(a) as being obvious over "Hierarchical Classification of Web Content", hereinafter referred to as Dumais et al., in view of "SMOTE: Synthetic Minority Over-sampling Technique", hereinafter referred to as Chawla et al., further in view of US 20070250501, hereinafter referred to as Grubb et al.

Regarding **claim 17**, Dumais et al., as modified by Chawla et al. and Rathod, discloses the method of claim 12, but not further comprising parsing the data to extract predetermined information related to the ~~deep~~ web forums including a topic title, a topic content, a post content, a topic author, a post author, an author reputation, and a posted date. Grubb et al. is cited to disclose parsing the data to extract predetermined information related to the ~~deep~~ web forums including a topic title, a topic content, a post content, a topic author, a post author, an author reputation, and a posted date ("All

users, for each item that is saved, can specify metadata about the items including, but not limited to: title, tags, categories, topics, keywords, date, URL, referring URL, rating, comments, quotations from the item, author, publication, date, source, ISBN or ISSN, library cataloging data, date stamps and/or bibliographic data," Grubb et al., [0032].). *Grubb et al. benefits Dumais et al. by tagging web content according to topic titles (Grubb et al., [0032]), thereby allowing a user to quickly ascertain a distribution of relevant topics. Therefore, it would be obvious to one skilled in the art to combine the teachings of Dumais et al. with those of Grubb et al. to improve the web content classification of Dumais et al.*

10. **Claims 18 and 19** are rejected under pre-AIA 35 U.S.C. 103(a) as being obvious over "Hierarchical Classification of Web Content", hereinafter referred to as Dumais et al., in view of US 20180246983, hereinafter referred to as Rathod.

Regarding **claim 18**, Dumais et al. discloses a method, comprising:

~~configuring a processor for executing operations including:~~

accessing data associated with a ~~deep~~ web forum, the data defining a topic for classification ("During the testing or operational phase, the learned classifier is used to classify new web pages," Dumais et al., p. 259, Text Classification Using SVMs section, 1st para.);

extracting a set of features from the data as inputs for a machine classifier (See Dumais et al., fig. 1. And, "After preprocessing, a binary vector was created for each page

indicating whether each term appeared in the page summary or not," Dumais et al., p.259, right col., 2nd para.); and

apply a machine classifier to the set of features to generate a prediction list of tags for classifying the topic ("A support vector machine (SVM) algorithm was used as the classifier, because it has been shown in previous work to be both very fast and effective for text classification problems," Dumais et al., p. 259, Text Classification Using SVMs section, 2nd para.), wherein the machine classifier is configured to maintain a predetermined tag hierarchy by regulating modification to parent and child tags associated with a predicted tag of the prediction list of tags based on a prediction probability of the predicted tag and predetermined tag thresholds assigned to the parent and child tags ("For the hierarchical models, p=0.2 for the multiplicative decision rule, and p_1=0.2 (first level) and p_2=0.5 (second level) for the Boolean decision rule," Dumais et al., p. 260, SVM Parameters section, 2nd para. Here, first level is the parent level/threshold corresponding to a first probability and second level is the child level/threshold corresponding to a second probability.).

Dumais et al., though, does not specifically disclose a *deep* web forum or a processor for performing the above steps.

Rathod is cited to disclose a *deep* web forum ("Internet content that is not capable of being searched by a web search engine is generally described as the deep web. Web search engine," Rathod, para [0326].) and a processor for performing the above steps (Rathod, para [0206]). *Rathod benefits Dumais et al. by expanding the document classification techniques described in Dumais to deep web forums (Rathod, para*

[0326]). Therefore, it would be obvious for one skilled in the art to combine the teachings of Dumais et al. with those of Rathod to extend the application of the hierarchical classification of Dumais to deep web content.

Regarding **claim 19**, Dumais et al., as modified by Rahod, discloses the method of claim 18, further comprising:

outputting an update to the prediction list of tags, a total number of the tags of the prediction list of tags being different from the update to the prediction list of tags to maintain the predetermined tag hierarchy ("FIGS. 9-10 discuss in detail about creation, updating, previewing and managing structured pages and categories 707," Rathod, para [0289].).

11. **Claim 20** is rejected under pre-AIA 35 U.S.C. 103(a) as being obvious over "Hierarchical Classification of Web Content", hereinafter referred to as Dumais et al., in view of US 20180246983, hereinafter referred to as Rathod, and further in view of US 20130097103, hereinafter referred to as Chari et al.

Regarding **claim 20**, Dumais et al., as modified by Rathod, discloses the method of claim 18, but not further comprising generating the predetermined tag hierarchy based on a ground truth dataset, and leveraging semi-supervised learning to address class imbalance and increase accuracy of the machine classifier. Chari et al. is cited to disclose generating the predetermined tag hierarchy based on a ground truth dataset, and leveraging semi-supervised learning to address class imbalance and increase

accuracy of the machine classifier ("Now presented is an overview of the process which provides a high level intuitive guide through the methodology 100 (FIG. 1) for obtaining balanced training sets. The present techniques provide a solution where there is an unlabeled data set with unknown class distribution, and the goal is to produce balanced labeled samples for training predictive models. If one assumes that the labels of the samples in the data set are known a priori, one can use over and under-sampling to draw a balanced sample set. See, for example, Liu et al., "Exploratory under-sampling for class-imbalance learning," IEEE Trans. On Sys. Man. And Cybernetics (2009) (hereinafter "Liu"); Chawla et al., "Smote: Synthetic minority over-sampling technique," Journal of Artificial Research (JAIR), vol. 16, pgs. 321-357 (2002) (hereinafter "Chawla"); Wu et al., "Data selection for speech recognition," in IEEE workshop on Automatic Speech Recognition and Understanding (ASRU) (2007) (hereinafter "Wu"), the contents of each of which are incorporated by reference herein. In practice, however, the class labels are not known and instead a series of approximations must be used to approach the results of this ideal solution. An iterative method is applied herein, where, in each iteration, the present method draws (selects) a batch of samples (B), and domain experts provide the labels of the selected samples. Information embedded in the labeled sample is used to group together data elements which are very similar to the labeled sample using semi-supervised clustering. The class distribution in the clusters can then be estimated and used to perform a biased sampling of clusters to obtain a diverse balanced sample. Within each cluster, a diverse sample is obtained by using a maximum entropy sampling. The sample obtained at each iteration is then labeled and used in subsequent iterations," Chari et al., para [0028].). *Chari et al.*

benefits Dumais et al. by performing a biased sampling of clusters to obtain a diverse balanced sample (Chari et al., para [0028]). Therefore, it would be obvious for one skilled in the art to combine the teachings of Dumais et al. with those of Chari et al. to improve the classification scheme of Dumais et al.

Conclusion

1. Other prior art is noted on attached PTO-892. In particular, the examiner notes Huillier et al. which, like Dumaias et al., described classification of topics on the Dark Web.

2. Any inquiry concerning this communication or earlier communications from the examiner should be directed to ANNE L THOMAS-HOMESCU whose telephone number is (571)272-0899. The examiner can normally be reached on Mon-Fri 8-6.

Examiner interviews are available via telephone, in-person, and video conferencing using a USPTO supplied web-based collaboration tool. To schedule an interview, applicant is encouraged to use the USPTO Automated Interview Request (AIR) at http://www.uspto.gov/interviewpractice.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Bhavesh M Mehta can be reached on 5712727453. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR.

Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see https://ppair-my.uspto.gov/pair/PrivatePair. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/ANNE L THOMAS-HOMESCU/

Primary Examiner, Art Unit 2656

Notice of References Cited	Application/Control No. 16/405,612	Applicant(s)/Patent Under Reexamination Patil et al.	
	Examiner ANNE L THOMAS-HOMESCU	Art Unit 2656	Page 1 of 4

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
*	A	US-20180121539-A1	05-2018	Ciulla; Fabio	G06F16/3344	1/1
*	B	US-20200151222-A1	05-2020	MANNAR; Kamal	G06K9/4628	1/1
*	C	US-20180082211-A1	03-2018	Allen; Corville O.	G06F16/9535	1/1
*	D	US-20140250032-A1	09-2014	Huang; Shu	G06N20/00	706/12
*	E	US-20200012673-A1	01-2020	RUDZICZ; Frank	G06K9/6264	1/1
*	F	US-20180329935-A1	11-2018	Mugali; Aditya	G06F16/282	1/1
*	G	US-20080162385-A1	07-2008	Madani; Omid	G06F16/353	706/12
*	H	US-20040064464-A1	04-2004	Forman, George Henry	G06F16/35	1/1
*	I	US-20130117451-A1	05-2013	DU; Lin	H04L63/102	709/225
*	J	US-20130132311-A1	05-2013	Liu; Jingchen	G06K9/6256	706/12
*	K	US-20180025059-A1	01-2018	Batchu; Ravi	G06F16/24564	707/740
*	L	US-20200180148-A1	06-2020	S NANAL; Mihir	G06F11/0793	1/1
*	M	US-20190180246-A1	06-2019	Ferguson; David	G06F40/20	1/1

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
*	U	Dumais et al., "Hierarchical Classification of Web Content", ACM 2000. (Year: 2000)
*	V	Dalin et al., "Criminal motivation on the dark web: A categorisation model for law enforcement", Digital Investigation, 2018. (Year: 2018)
*	W	Hassan et al., "A Task Level Metric for Measuring Web Search Satisfaction and its Application on Improving Relevance Estimation", ACM, 2011. (Year: 2011)
*	X	Chawla, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, 2002. (Year: 2002)

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
*	A	US-20170193060-A1	07-2017	Khanzode; Sandeep	G06F16/338	1/1
*	B	US-20090307196-A1	12-2009	Shuster; Gary Stephen	H04L67/22	1/1
*	C	US-20190347359-A1	11-2019	Guy; Ido	G06F16/9574	1/1
*	D	US-20080159495-A1	07-2008	Dahan; Jean-David	G10L25/51	379/93.01
*	E	US-8972376-B1	03-2015	Gailis; Renars	G06F16/951	707/710
*	F	US-20150026159-A1	01-2015	Berkman; Paul Arthur	G06F16/9535	707/722
*	G	US-20110258560-A1	10-2011	Mercuri; Marc E.	G06F16/958	715/753
*	H	US-20100281022-A1	11-2010	Gollapudi; Sreenivas	G06F16/9024	707/723
*	I	US-20080189316-A1	08-2008	Otsuka; Nobuo	G06F16/3335	1/1
*	J	US-20060206476-A1	09-2006	Kapur; Shyam	G06F16/30	1/1
*	K	US-20020078045-A1	06-2002	Dutta, Rabindranath	G06F16/951	1/1
*	L	US-20070250501-A1	10-2007	Grubb; Michael L.	G06F16/3343	1/1
*	M	US-20170046180-A1	02-2017	DESINENI; Kalyan	G06F9/542	1/1

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Notice of References Cited	Application/Control No. 16/405,612	Applicant(s)/Patent Under Reexamination Patil et al.	
	Examiner ANNE L THOMAS-HOMESCU	Art Unit 2656	Page 3 of 4

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
*	A	US-20190236412-A1	08-2019	ZHAO; Wei	G06K9/6256	1/1
*	B	US-20190230070-A1	07-2019	Isaacson; Thomas M.	G07G1/0036	1/1
*	C	US-20190205885-A1	07-2019	Lim; Chern Jie	G06Q20/045	1/1
*	D	US-20200159225-A1	05-2020	Zeng; Wenyuan	G05D1/0088	1/1
*	E	US-20190133533-A1	05-2019	ALAM; Shahnawaz	A61B5/7221	1/1
*	F	US-20170032276-A1	02-2017	SUKHANOV; Sergey	G06F16/41	1/1
*	G	US-20130097103-A1	04-2013	Chari; Suresh N.	G06N20/00	706/12
*	H	US-20190362069-A1	11-2019	Park; Jason	G06F16/904	1/1
*	I	US-20180246983-A1	08-2018	Rathod; Yogesh	G06F16/9535	1/1
	J					
	K					
	L					
	M					

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Notice of References Cited	Application/Control No. 16/405,612	Applicant(s)/Patent Under Reexamination Patil et al.	
	Examiner ANNE L THOMAS-HOMESCU	Art Unit 2656	Page 4 of 4

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
	A					
	B					
	C					
	D					
	E					
	F					
	G					
	H					
	I					
	J					
	K					
	L					
	M					

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	L'Huillier, "Topic-Based Social Network Analysis for Virtual Communities of Interests in the Dark Web", SIGKDD Explorations, 2010. (Year: 2010)
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	16/405,612	Patil et al.
	Examiner	Art Unit
	ANNE L THOMAS-HOMESCU	2656

CPC - Searched*		
Symbol	Date	Examiner
G06F40/30,221,284 16/951; G06N20/00,10,20, 5/003, 3/084	02/19/2021	ATH

CPC Combination Sets - Searched*		
Symbol	Date	Examiner

US Classification - Searched*			
Class	Subclass	Date	Examiner

* See search history printout included with this form or the SEARCH NOTES box below to determine the scope of the search.

Search Notes		
Search Notes	Date	Examiner
See attached EAST and Google Scholar searches.	02/19/2021	ATH

Interference Search			
US Class/CPC Symbol	US Subclass/CPC Group	Date	Examiner

/ANNE L THOMAS-HOMESCU/ Primary Examiner, Art Unit 2656	

IPR2025-01398
CrowdStrike EX1002 Page 249

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
	16/405,612	Patil et al.
	Examiner	Art Unit
	ANNE L THOMAS-HOMESCU	2656

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

CLAIMS

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	02/17/2021								
	1	✓								
	2	✓								
	3	✓								
	4	✓								
	5	✓								
	6	✓								
	7	✓								
	8	✓								
	9	✓								
	10	✓								
	11	✓								
	12	✓								
	13	✓								
	14	✓								
	15	✓								
	16	✓								
	17	✓								
	18	✓								
	19	✓								
	20	✓								

EAST Search History

EAST Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
S60	4	("10176438").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:33
S61	7	("10313385").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:36
S62	2	("10437945").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:37
S63	2	("20190349393").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:38
S64	3	("20200036743").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:39
S65	5	("2019089389").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:40

S66	4	("2019157335").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:41
S67	2,734	(deep near3 web)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:47
S68	3,521	((deep dark) near3 web)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:48
S69	4,198	((deep dark) near3 (web internet))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:51
S70	8,632	((deep dark) near7 (web internet))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:51
S71	2,196	((deep dark) near3 (web internet)) and (forum$1 discuss$4 blog$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:53
S72	34	((deep dark) near3 (web internet)) and (forum$1 discuss$4 blog$4 meeting$1 conference$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:56

S73	34	((deep dark) near3 (web internet)) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:57
S74	34	((deep dark) near3 (web internet) D2web) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:59
S75	50	((deep dark) near3 (web internet) D2web) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 10:02
S76	11,294	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 10:05
S77	6	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (parent child hierarch$3) same (train$3 learn$3 adapt$4 adjust$4) and (criminal$3 illegal$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 10:57
S78	32	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (parent child hierarch$3) same (train$3 learn$3 adapt$4 adjust$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 10:58
S79	6	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media chat$4) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (parent child hierarch$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 10:59

		same (train$3 learn$3 adapt$4 adjust$4) and (criminal$3 illegal$4)				
S80	5,248	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) and (index$4 (dark near3(web internet cyber)))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:08
S81	886	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) and (index$4 (dark near3(web internet cyber)))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:11
S82	31	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child) and (index$4 (dark near3(web internet cyber)))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:29
S83	107	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) and (index$4 (dark near3 (web internet cyber)))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:30
S84	76	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) and (index$4 (dark near3 (web internet cyber))) NOT S82	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:45

S85	2	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) and (deep near3 web) NOT S82	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:53
S86	2	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child) and (deep near3 web)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:56
S87	144	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:56
S88	50	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:58
S89	37	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:59

S90	50	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 13:05
S91	52	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media post$3) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 13:12
S92	238	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media post$3) same (topic$1 subject$1 title$1 theme$1 classification classif$4 class$2 categor$3 categori$1ation group$4) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 13:15
S93	1,797	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media post$3) same (topic$1 subject$1 title$1 theme$1 classification classif$4 class$2 categor$3 categori$1ation group$4) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 13:16
S94	110	(dark$1net ((dark deep) near3 web)) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO;	OR	OFF	2021/02/16 13:34

			DERWENT; IBM_TDB			
S95	1,005	(dark$1net ((dark deep) near3 (web internet cyber network$1))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 13:55
S96	895	(dark$1net ((dark deep) near3 (web internet cyber network$1))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged) NOT S94	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 13:58
S97	31	(dark$1net ((dark deep) near3 (web internet cyber network$1))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) NOT S94	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 14:02
S98	0	(dark$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) NOT S94	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 14:05
S99	53	(dark$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged identifier$1 identification$1 identif$4) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1) NOT S94	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 14:08
S100	55	(dark$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged identifier$1 identification$1 identif$4 cluster$3 group$4) same (forum$1 discuss$4 blog$4 meeting$1	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 14:24

		conference$1 panel$4 media page$1 site$1 web$1site$1) NOT S94				
S101	46	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 16:48
S102	166	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 16:57
S103	1,092	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 17:18
S104	36	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 17:22

		confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec) and (ground near3 truth)				
S105	941	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec) and ((ground near3 truth) training)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 17:52
S106	7	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (hierarch$3 sub$1categor$3 sub$1class$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec) and (ground near3 truth)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 18:03
S107	48	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (hierarch$3 sub$1categor$3 sub$1class$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec) and ((ground near3 truth) training)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 18:04
S108	52	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1	US-PGPUB; USPAT;	OR	OFF	2021/02/16 18:07

		conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (hierarch$3 sub$1categor$3 sub$1class$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec)	USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB			
S109	77	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (hierarch$3 sub$1categor$3 sub$1class$3 parent child) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 18:10
S110	2	("7349917").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 18:25
S111	110	(dark$1net ((dark deep) near3 web)) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 13:22
S112	51	(dark$1net crime$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged group$3) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 13:22
S113	0	(dark$1net crime$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3)	US-PGPUB; USPAT; USOCR;	OR	OFF	2021/02/17 13:25

		same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged group$3) same (title$1) and (process$3) NOT S111	FPRS; EPO; JPO; DERWENT; IBM_TDB			
S114	51	(dark$1net crime$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged group$3) and (process$3) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 13:26
S115	10	(dark$1net crime$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 13:32
S116	48	(dark$1net crime$1net ((dark deep) near3 (web internet cyber))) and (crime$1 criminal$3 illegal$3 fraud faudulent$3 theft$1) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 13:33
S117	0	(dark$1net crime$1net ((dark deep) near3 (web internet cyber))) and (crime$1 criminal$3 illegal$3 fraud faudulent$3 theft$1) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 title$1) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 14:37
S118	63,544	(classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 title$1) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 14:44
S119	3,622	(classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 subject$1) same (title$1) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 14:45
S120	1,690	(web$1site$1 webpage$1) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged	US-PGPUB; USPAT; USOCR;	OR	OFF	2021/02/17 15:25

		cluster$3) near7 (topic$1 subject$1) same (title$1) NOT S111	FPRS; EPO; JPO; DERWENT; IBM_TDB			
S121	575	(web$1site$1 webpage$1) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 subject$1) near7 (title$1) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 15:25
S122	40	(web$1site$1 webpage$1) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 subject$1) near7 (title$1) same (vector$1 feature$1 word2vec) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 15:26
S123	0	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 16:27
S124	4	(web internet cyber$4) and (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 16:28
S125	435	(web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 08:47

S126	1	(web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 08:50
S127	1	(web internet cyber$4) and (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 08:51
S128	1	(probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 08:51
S129	1	(web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3 metric$4) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 09:13
S130	1,030	(web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same (character$1 word$1) near7 (remov$3 delet$3 ignor$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:19
S131	855	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same (character$1 word$1) near7 (remov$3 delet$3 ignor$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:25

S132	118	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same (character$1) same (word$1) near7 (remov$3 delet$3 ignor$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:26
S133	315	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same (character$1) same (word$1) same (remov$3 delet$3 ignor$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:27
S134	50	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same ((character$1) same (word$1)) near7 (remov$3 delet$3 ignor$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:31
S135	1	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same ((character$1) same (word$1)) near7 (remov$3 delet$3 ignor$4) and (root$1) same (stem$4) same (lemmati$1ation lemma$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:35
S136	1	(G10L.cpcl. G06F.cpcl.) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same ((character$1) same (word$1)) near7 (remov$3 delet$3 ignor$4) and (root$1) same (stem$4) same (lemmati$1ation lemma$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:37
S137	131	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2	US-PGPUB; USPAT; USOCR;	OR	OFF	2021/02/18 10:38

		classification categori$1ation categori$4 group$3 cluster$3) and (root$1) same (stem$4) same (lemmati$1ation lemma$1)	FPRS; EPO; JPO; DERWENT; IBM_TDB			
S138	28,160	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (crawl$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:48
S139	6,829	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (crawl$3) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:49
S140	14,864	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (crawl$3) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media site$1 web$1site$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:49
S141	5,259	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (crawl$3) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (internet web site$1 web$1site$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:51
S142	2,342	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (crawler) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (internet web site$1 web$1site$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:53
S143	838	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (crawler) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (internet web site$1 web$1site$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:54

S144	354	(G10L.cpcl. G06F.cpcl.) and (classif$4 class$2 classification categori$1ation categori$4) same (crawler) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (internet web site$1 web$1site$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:57
S145	56	(G10L.cpcl. G06F.cpcl.) and (classif$4 class$2 classification categori$1ation categori$4) same (crawl$3) same (spider$4) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (internet web site$1 web$1site$1 html)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 12:25
S146	110	(dark$1net ((dark deep) near3 web)) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 13:29
S147	0	(web$1site$1 webpage$1) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 subject$1) near7 (title$1) same (date$1) same (reputation$1 review$4 comment$4) same (author$1 writer$1) same (vector$1 feature$1 word2vec) NOT S146	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 13:30
S148	7	(web$1site$1 webpage$1) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 subject$1) near7 (title$1) same (date$1) same (reputation$1 review$4 comment$4) same (author$1 writer$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 13:32
S149	20	(web$1site$1 webpage$1) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) same (topic$1 subject$1) near7 (title$1) same (date$1) same (reputation$1 review$4 comment$4) same (author$1 writer$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 13:36
S150	3	(G10L.cpcl. G06F.cpcl.) and (classif$4 class$2 classification categori$1ation categori$4) same (crawl$3) same (spider$4) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (internet web	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO;	OR	OFF	2021/02/18 13:56

		site$1 web$1site$1 html) same (train$3 (ground near3 truth))	DERWENT; IBM_TDB			
S151	1,784	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 16:23
S152	1,785	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 16:24
S153	273	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (semi$1supervis$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 16:26
S154	1	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (semi$1supervis$4) same (synthetic$4 synthes$4 artificial$4) near7 (sampl$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 17:57
S155	31	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (nearest near3 neighbor$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 18:50
S156	1	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (nearest near3 neighbor$1) same (elastic$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 21:08
S158	4	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation	US-PGPUB; USPAT;	OR	OFF	2021/02/18 21:13

		categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4)	USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB			
S159	4	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4 BM25)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 21:33
S160	25	((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4 BM25)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 21:35
S161	3,029	(imbalance$1 minorit$3) same (elastic$4 BM25)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 21:44
S162	303	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (imbalance$1 minorit$3) same (elastic$4 BM25)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 21:44
S163	3	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (imbalance$1 minorit$3) same (elastic$4 BM25)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 21:45
S164	108	(dark$1net ((dark deep) near3 web)) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged) and (process$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/19 09:07
S165	1,898	(dark$1net ((dark deep) near3 web)) and (classification classif$4	US-PGPUB; USPAT; USOCR;	OR	OFF	2021/02/19 09:20

		categori$1ation categor$3 label$4 tag tagging tagged) and (process$3)	FPRS; EPO; JPO; DERWENT; IBM_TDB			
S166	287	(dark$1net ((dark deep) near3 web)) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged) near7 (updat$3 adjust$4 adaptation adapt$3) and (process$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/19 09:59
S167	292	(dark$1net ((dark deep) near3 web)) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged) near7 (updat$3 adjust$4 adaptation adapt$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/19 10:30

EAST Search History (Interference)

<This search history is empty>

2/19/2021 2:12:55 PM
C:\Users\athomashomescu\Documents\EAST\Workspaces\16405612 - assigning labels to discussion topics shown in online forums on the dark web.wsp

INFORMATION DISCLOSURE STATEMENT BY APPLICANT

(Not for submission under 37 CFR 1.99)

Application Number	16405612
Filing Date	2019-05-07
First Named Inventor	Revanth Patil
Art Unit	2658
Examiner Name	Richemond Dorvil
Attorney Docket Number	055743-621457 M18-191P

U.S.PATENTS

Examiner Initial*	Cite No	Patent Number	Kind Code[1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1	10176438	B2	2019-01-08	Shakarian et al.	
	2	10313385	B2	2019-06-04	Shakarian et al.	
	3	10437945	B2	2019-10-08	Shakarian et al.	

If you wish to add additional U.S. Patent citation information please click the Add button.

U.S.PATENT APPLICATION PUBLICATIONS

Examiner Initial*	Cite No	Publication Number	Kind Code[1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1	20190349393	A1	2019-11-14	Nunes et al.	
	2	20200036743	A1	2020-01-30	Almukaynizi et al.	

If you wish to add additional U.S. Published Application citation information please click the Add button.

FOREIGN PATENT DOCUMENTS

Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2]	Kind Code[4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear	T[5]

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /A.L.T/

IPR2025-01398
CrowdStrike EX1002 Page 270

Receipt date: 08/17/2020

INFORMATION DISCLOSURE STATEMENT BY APPLICANT

(Not for submission under 37 CFR 1.99)

Application Number	16405612 16/405,612 - GAU: 2656
Filing Date	2019-05-07
First Named Inventor	Revanth Patil
Art Unit	2658
Examiner Name	Richemond Dorvil
Attorney Docket Number	055743-621457 M18-191P

	1	2019089389	WO	A1	2019-05-09	Shakarian et al.		☐
	2	2019157335	WO	A1	2019-08-15	Shaabani et al.		☐

If you wish to add additional Foreign Patent Document citation information please click the Add button

NON-PATENT LITERATURE DOCUMENTS

Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc), date, pages(s), volume-issue number(s), publisher, city and/or country where published.	T5
	1	U.S. Application No. 16/424,929, Shakarian et al., filed 2016-11-28.	☐
	2	U.S. Application No. 16/548,329, Tavabi et al., filed 2019-08-22.	☐
	3	U.S. Application No. 16/640,878, Shakarian et al., filed 2018-10-26.	☐
	4	U.S. Application No. 16/653,899, Sarkar et al., filed 2019-10-15.	☐
	5	U.S. Application No. 16/875,540, Nunes et al., filed 2020-05-15.	☐
	6	Tor Project, "Tor: Overview" [Online] Available: https://www.torproject.org/about/overview.html.en , accessed 27-December-2018" ("2017")	☐
	7	Abbasi, et al., "Descriptive analytics: Examining expert hackers in web forums", Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint, pp. 56–63 (IEEE, 2014).	☐

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /A.L.T/

Receipt date: 08/17/2020

INFORMATION DISCLOSURE STATEMENT BY APPLICANT

(Not for submission under 37 CFR 1.99)

Application Number	16405612 16/405,612 - GAU: 2656
Filing Date	2019-05-07
First Named Inventor	Revanth Patil
Art Unit	2658
Examiner Name	Richemond Dorvil
Attorney Docket Number	055743-621457 M18-191P

	8	Breiman, et al., "Random forests-classification description", Department of Statistics, Berkeley 2 (2007).	☐
	9	Chawla, et al., "Smote: synthetic minority over-sampling technique", Journal of artificial intelligence research 16, 321–357 (2002).	☐
	10	Chertoff, et al., "The impact of the dark web on internet governance and cyber security", (2015).	☐
	11	Dekel, et al., "Multiclass-multilabel classification with more classes than examples", Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 137–144 (2010).	☐
	12	Harris, "Distributional structure", Word 10, 2-3, 146–162 (1954).	☐
	13	Le, et al., "Distributed representations of sentences and documents", Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 1188–1196 (2014).	☐
	14	Manning, "Introduction", in Retrieving the Radical Tillich, pp. 1–19 (Springer, 2015).	☐
	15	Marin, et al., "Product offerings in malicious hacker markets", Intelligence and Security Informatics (ISI), 2016 IEEE Conference, pp. 187– 189 (IEEE, 2016).	☐
	16	Moore, et al., "Cryptopolitik and the darknet", Survival 58, 1, 7–38 (2016).	☐
	17	Nunes, et al., "Darknet and deepnet mining for proactive cybersecurity threat intelligence", Intelligence and Security Informatics (ISI), 2016 IEEE Conference, pp. 7–12 (IEEE, 2016).	☐
	18	Okapi-BM25, "Okapi-bm25 ranking function Wikipedia, the free encyclopedia", URL https://en.wikipedia.org/wiki/Okapi-BM25, [Online; accessed 12-December-2017] (2017).	☐

IPR2025-01398
CrowdStrike EX1002 Page 272

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /A.L.T/

Receipt date: 08/17/2020

INFORMATION DISCLOSURE STATEMENT BY APPLICANT

(Not for submission under 37 CFR 1.99)

Application Number	16405612 16/405,612 - GAU: 2656
Filing Date	2019-05-07
First Named Inventor	Revanth Patil
Art Unit	2658
Examiner Name	Richemond Dorvil
Attorney Docket Number	055743-621457 M18-191P

	19	Pedregosa, et al., "Scikit-learn: Machine learning in Python", Journal of Machine Learning Research 12, 2825–2830 (2011).	☐
	20	Pérez-Iglesias, et al., "Integrating the probabilistic models bm25/bm25f into lucene", arXiv preprint arXiv:0911.5046 (2009).	☐
	21	Rajaraman, Jd (2011). Data mining, Mining of Massive Datasets pp. 1–17 (2011).	☐
	22	Rong, "word2vec parameter learning explained", arXiv preprint arXiv:1411.2738 (2014).	☐
	23	Samtani, et al., "Azsecure hacker assets portal: Cyber threat intelligence and malware analysis", Intelligence and Security Informatics (ISI), 2016 IEEE Conference, pp. 19–24 (IEEE, 2016).	☐
	24	Sanderson, Introduction to information retrieval, cambridge university press 2009, Natural Language Engineering 16, 1, 232–234 (2010).	☐
	25	Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval", Journal of documentation 28, 1, 11–21 (1972).	☐
	26	Wikipedia, "Random forest — wikipedia, the free encyclopedia", URL https://en.wikipedia.org/w/index.php?title=Randomforestoldid= 822509370, [Online; accessed29 − December − 2018].	☐

If you wish to add additional non-patent literature document citation information please click the Add button

EXAMINER SIGNATURE

Examiner Signature	/ANNE L THOMAS-HOMESCU/	Date Considered	02/16/2021

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document. [4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /A.L.T/

Receipt date: 08/17/2020

INFORMATION DISCLOSURE STATEMENT BY APPLICANT

(Not for submission under 37 CFR 1.99)

Application Number	16405612 16/405,612 - GAU: 2656
Filing Date	2019-05-07
First Named Inventor	Revanth Patil
Art Unit	2658
Examiner Name	Richemond Dorvil
Attorney Docket Number	055743-621457 M18-191P

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

☐ That each item of information contained in the information disclosure statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☐ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

☐ See attached certification statement.

☐ The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

☒ A certification statement is not submitted herewith.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	/Ari M. Bai/	Date (YYYY-MM-DD)	2020-08-17
Name/Print	Ari M. Bai	Registration Number	38726

This collection of information is required by 37 CFR 1.97 and 1.98. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 1 hour to complete, including gathering, preparing and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /A.L.T/

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these record s.

2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.

3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.

4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).

5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.

6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).

7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.

8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.

9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

IPR2025-01398

Bibliographic Data

Application No: 16/405,612

Foreign Priority claimed: ○ Yes ◉ No

35 USC 119 (a-d) conditions met: ☐ Yes ☑ No ☐ Met After Allowance

Verified and Acknowledged: /ANNE L THOMAS-HOMESCU/ (Examiner's Signature) ATH (Initials)

Title: SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB

FILING or 371(c) DATE	CLASS	GROUP ART UNIT	ATTORNEY DOCKET NO.
05/07/2019	704	2656	055743-621457 (M18-191P)
RULE			

APPLICANTS

INVENTORS

Revanth Patil, Tempe, AZ, UNITED STATES

Paulo Shakarian, Tempe, AZ, UNITED STATES

Ashkan Aleali, Tempe, AZ, UNITED STATES

Ericsson Marin, Tempe, AZ, UNITED STATES

CONTINUING DATA

This application has PRO of 62668878 05/09/2018

FOREIGN APPLICATIONS

IF REQUIRED, FOREIGN LICENSE GRANTED**

05/17/2019

** MICRO ENTITY **

STATE OR COUNTRY

UNITED STATES

ADDRESS

POLSINELLI PC
900 WEST 48TH PLACE
SUITE 900
KANSAS CITY, MO 64112-1895
UNITED STATES

FILING FEE RECEIVED

$470

PTO/SB/06 (09-11)
Approved for use through 1/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 16/405,612	Filing Date 05/07/2019	☐ To be Mailed

ENTITY: ☐ LARGE ☐ SMALL ☑ MICRO

APPLICATION AS FILED - PART I

FOR	(Column 1) NUMBER FILED	(Column 2) NUMBER EXTRA		RATE ($)	FEE ($)
☑ BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A		N/A	75
☑ SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A		N/A	165
☑ EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A		N/A	190
TOTAL CLAIMS (37 CFR 1.16(i))	minus 20 =	*		x $25 =	
INDEPENDENT CLAIMS (37 CFR 1.16(h))	minus 3 =	*		x $115 =	
☐ APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $310 ($155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).				
☐ MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))					
* If the difference in column 1 is less than zero, enter "0" in column 2.				TOTAL	430

APPLICATION AS AMENDED - PART II

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT	08/13/2021	CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	* 20	Minus	** 20	= 0		x $25 =	0
	Independent (37 CFR 1.16(h))	* 3	Minus	*** 3	= 0		x $120 =	0
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	0

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	*	Minus	**	=		x $0 =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=		x $0 =	
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.	LIE
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".	/KATINA R TOBIN/
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".	
The "Highest Number Previously Paid For" (Total or Independent) is the highest number found in the appropriate box in column 1.	

This collection of information is required by 37 CFR 1.16. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s):	Revanth Patil	Examiner:	Anne L. Thomas-Homescu
Serial No.:	16/405,612	Group Art Unit:	2656
Filed:	May 7, 2019	Confirmation No.:	4562

For: SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARKWEB

AMENDMENT & RESPONSE UNDER 37 C.F.R. §1.111

TO THE COMMISSIONER FOR PATENTS

Sir:

This paper is responsive to the non-final Office Action mailed on February 24, 2021. Please enter the following amendments and consider the remarks below in accordance with the Rules of Practice. A three-month extension of time including the fee for extending the deadline to August 24, 2021 is included.

No further fees are believed to be due in connection with the filing of this paper; however, should any fees be deemed necessary, the commissioner is hereby authorized to deduct any such fees from **Deposit Account No. 50-1662** referencing the above docket number.

Amendments to the claims are reflected in the listing that begins at page 2 of this paper.

Remarks begin at page 7 of this paper.

Electronic Patent Application Fee Transmittal

Application Number:	16405612
Filing Date:	07-May-2019
Title of Invention:	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB
First Named Inventor/Applicant Name:	Revanth Patil
Filer:	Timothy D. Fontes/Rebecca Endsley
Attorney Docket Number:	055743-621457 (M18-191P)

Filed as Micro Entity

Filing Fees for Utility under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Extension - 3 months with $0 paid	3253	1	370	370
Miscellaneous:				
			Total in USD ($)	**370**

Electronic Acknowledgement Receipt

EFS ID:	43510921
Application Number:	16405612
International Application Number:	
Confirmation Number:	4562
Title of Invention:	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB
First Named Inventor/Applicant Name:	Revanth Patil
Customer Number:	27148
Filer:	Timothy D. Fontes/Rebecca Endsley
Filer Authorized By:	Timothy D. Fontes
Attorney Docket Number:	055743-621457 (M18-191P)
Receipt Date:	13-AUG-2021
Filing Date:	07-MAY-2019
Time Stamp:	18:48:22
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	DA
Payment was successfully received in RAM	$370
RAM confirmation Number	E20218CI48364936
Deposit Account	
Authorized User	

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		055743_621457_Response_to_OA.pdf	231995 6122e278928e9ebbc82ad1e0328d2eb7d1784f73	yes	12

	Multipart Description/PDF files in .zip description		
	Document Description	**Start**	**End**
	Amendment/Req. Reconsideration-After Non-Final Reject	1	1
	Claims	2	6
	Applicant Arguments/Remarks Made in an Amendment	7	12

Warnings:					
Information:					
2	Fee Worksheet (SB06)	fee-info.pdf	38748 aaa544761eb915a457e080d611f50d766983d29d	no	2
Warnings:					
Information:					
			Total Files Size (in bytes):	270743	

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

CLAIM AMENDMENTS

Please amend the claims as follows:

1. (Currently Amended) A computer-implemented system for improving classification of criminal activities arising from cyber environments, comprising:

a processor configured to:

access a ground truth dataset generated from deep web forum information, the ground truth dataset defining a predetermined tag hierarchy including predetermined tag labels corresponding topics of the deep web forum information;

access data related to a discussion topic for classification from a deep web forum;

extract a set of features from the data by assigning word vectors or paragraph vectors to the data; and

apply a machine classifier to the set of features to generate a prediction list of tags for classifying the discussion topic, wherein the prediction list includes a prediction probability value for each tag of the plurality of tags; and

add all parent tags associated with a tag of the plurality of tags to the prediction list based on a comparison between the prediction probability value for the tag and a first predetermined threshold value.

~~wherein the machine classifier is trained to maintain the predetermined tag hierarchy by regulating modification to parent and child tags associated with a predicted tag of the prediction list of tags based on a prediction probability of the predicted tag and predetermined tag thresholds assigned to the parent and child tags.~~

2. (Currently Amended) The computer-implemented system of claim 1, wherein ~~the~~ a predetermined tag hierarchy is divided into parent tags and child tags, wherein

a topic represented by a child tag is representative of a subset of a topic represented by a parent tag, wherein each child tag corresponds to a particular parent tag, and each tag is assigned a corresponding probability value.

3. (Currently Amended) The computer-implemented system of claim 1, wherein the processor is configured to remove all ~~a~~ child tags of the plurality of tags ~~of the prediction list of tags~~ corresponding to a parent tag from the prediction list of tags where a probability value associated with the child tag is below a second predetermined threshold value.

4. (Currently Amended) The computer-implemented system of claim 1, wherein the processor is configured to add all ~~a~~ parent tags of the plurality of tags corresponding to a child tag if a probability value associated with the parent tag is above the first ~~a~~ predetermined threshold value.

5. (Original) The computer-implemented system of claim 1, wherein the set of features are extracted using textual feature extraction to identify tags associated with topic titles of the data.

6. (Original) The computer-implemented system of claim 1, wherein a set of synthetic data samples are created for addition to the ground truth dataset to address class imbalance by creating a set of synthetic sample points along a line defined between a first minority feature vector and a second feature vector and adding the set of synthetic sample points to the ground truth dataset.

7. (Original) The computer-implemented system of claim 1, wherein the ground truth dataset is supplemented by, for every sample of the ground truth dataset in a minority class, adding top similar documents from the deep web forum information, as identified using an elastic search similarity score, and adding them to the ground truth dataset.

8. (Original) The computer-implemented system of claim 1, wherein the data is parsed and segmented into tokens, and predetermined characters and predetermined common words are removed from the data.

9. (Original) The computer-implemented system of claim 1, wherein words within the data are reduced to their root form using stemming and lemmatization procedures.

10. (Original) The computer-implemented system of claim 1, wherein the set of features are extracted from the data using a neural network, wherein the neural network is configured to learn a continuous distributed vector representation for documents and words within the data.

11. (Original) The computer-implemented system of claim 1, further comprising:

a crawler in operable communication with the processor and operable to traverse a plurality of deep web websites to retrieve Internet documents forming at least a portion of the deep web forum information.

12. (Canceled)

13. (Canceled)

14. (Canceled)

15. (Canceled)

16. (Canceled)

17. (Canceled)

18. (Currently Amended) A method, comprising:

configuring a processor for executing operations including:

accessing data associated with a deep web forum, the data defining a topic for classification;

extracting a set of features from the data as inputs for a machine classifier; and

apply a machine classifier to the set of features to generate a prediction list of tags for classifying the topic, wherein the prediction list includes a prediction probability value for each tag of the plurality of tags; and

adding all parent tags associated with a tag of the plurality of tags to the prediction list based on a comparison between the prediction probability value for the tag and a first predetermined threshold value.

~~wherein the machine classifier is configured to maintain a predetermined tag hierarchy by regulating modification to parent and child tags associated with a predicted tag of the prediction list of tags based on a prediction probability of the predicted tag and predetermined tag thresholds assigned to the parent and child tags.~~

19. (Currently Amended) The method of claim 18, further comprising:

outputting an update to the prediction list of tags, a total number of the tags of the prediction list of tags being different from the update to the prediction list of tags to maintain a ~~the~~ predetermined tag hierarchy.

20. (Currently Amended) The method of claim 18, further comprising generating a ~~the~~ predetermined tag hierarchy based on a ground truth dataset, and leveraging semi-supervised learning to address class imbalance and increase accuracy of the machine classifier.

REMARKS

This paper responds to the Examiner's Non-Final Office Action mailed on February 24, 2021. Reconsideration and continued examination in view of the foregoing response is respectfully requested.

Status of Claims

Claims 1-20 were originally filed in the instant application, with claims 1, 12, and 18 being independent. By the present document, claims 1-11 and 18-20 are pending. Claims 1-4 and 18-20 have been amended for discussion with the Examiner to possibly overcome the rejections to the recent Office action dated February 24, 2021. Claims 12-17 have been cancelled.

§103 Rejections (Dumais, Dalins, Otsuka, Grubb, Hassan, Khanzode, Chawla, Otsuka, Rathod, Chari)

The Examiner rejected claims 1-4 and 10-11 under pre-AIA 35 U.S.C. §103(a) as being obvious over the article titled Hierarchical Classification of Web Content by Dumais et al. ("Dumais") in view of the article titled Criminal Motivation on the Dark Web: A Categorization Model for Law Enforcement by Dalins et al. ("Dalins") and further in view of U.S. Patent Application Publication No. 2008/0189316 to Otsuka ("Otsuka"). The Examiner further rejected claim 5 under pre-AIA 35 U.S.C. §103(a) as being obvious over Dumais in view of Dalins, further in view of Otsuka, and further in view of U.S. Patent Application Publication No. 2007/0250501 to Grubb et al. ("Grubb"). The Examiner rejected claim 6 under pre-AIA 35 U.S.C. §103(a) as being obvious over Dumais in view of Dalins, further in view of Otsuka, and further in view of Grubb. The Examiner further rejected claim 7 under pre-AIA 35 U.S.C. §103(a) as being obvious over Dumais in view of Dalins, further in view of Otsuka, and further in view of the article titled A Task Level Metric for Measuring Web Search Satisfaction and Its Application on Improving Relevance Estimation by Hassan et al. ("Hassan"). The Examiner further rejected claims 8 and 9 under pre-AIA 35 U.S.C. §103(a) as being obvious over Dumais in view of Dalins, further in view of Otsuka, and further in view of U.S. Patent Application

Publication No. 2008/0189316 to Khanzode et al. ("Khanzode"). The Examiner further rejected claims 12, 14, and 15 under pre-AIA 35 U.S.C. §103(a) as being obvious over Dumais in view of the article titled SMOTE: Synthetic Minority Over-sampling Technique by Chawla et al. ("Chawla"), and further in view of U.S. Patent Application Publication No. 2018/0246983 to Rathod ("Rathod"). The Examiner rejected claim 13 under pre-AIA 35 U.S.C. §103(a) as being obvious over Dumais in view of Chawla, further in view of Rathod, and further in view of Hassan. The Examiner rejected claim 16 under pre-AIA 35 U.S.C. §103(a) as being obvious over Dumais in view of Chawla, further in view of Rathod, and further in view of Otsuka. The Examiner rejected claim 16 under pre-AIA 35 U.S.C. §103(a) as being obvious over Dumais in view of Chawla, and further in view of Grubb. The Examiner rejected claims 18 and 19 under pre-AIA 35 U.S.C. §103(a) as being obvious over Dumais in view of Rathod. The Examiner further rejected claim 20 under pre-AIA 35 U.S.C. §103(a) as being obvious over Dumais in view or Rathod, and further in view of U.S. Patent Application Publication No. 2013/0097103 to Chari et al. ("Chari"). Applicants respectfully traverse these rejections.

Independent claim 1 as presently amended recites the following features (emphasis added in bold):

A computer-implemented system for improving classification of criminal activities arising from cyber environments, comprising:

a processor configured to:

access a ground truth dataset generated from deep web forum information, the ground truth dataset defining a predetermined tag hierarchy including predetermined tag labels corresponding topics of the deep web forum information;

access data related to a discussion topic for classification from a deep web forum;

extract a set of features from the data by assigning word vectors or paragraph vectors to the data; and

apply a machine classifier to the set of features to generate a prediction list of tags for classifying the discussion topic, wherein the prediction list includes a prediction probability value for each tag of the plurality of tags; and

add all parent tags associated with a tag of the plurality of tags to the prediction list based on a comparison between the prediction probability value for the tag and a first predetermined threshold value.

The features of claim 1 as amended are supported at least by [0048]-[0054] of the present application. The present amendments clarify features of the claims and emphasize that the system **adds all parent tags associated with a tag of the plurality of tags to the prediction list** based on a comparison between the prediction probability value for the tag and a first predetermined threshold value. The additional independent claim recites the same or similar features.

The Office Action states that Dumais discloses "wherein the machine classifier is trained to maintain the predetermined tag hierarchy by regulating modification to parent and child tags associated with a predicted tag of the prediction list of tags based on a prediction probability of the predicted tag and predetermined tag thresholds assigned to the parent and child tags". Specifically, the Office Action refers to p. 260, SVM Parameters section, 2nd paragraph of Dumais et al. that discusses two methods of combining hierarchical results using a Boolean approach and a multiplicative approach. Additional discussion about the methods of combining hierarchical results can be found in p. 261, *Second Level, Hierarchical*, first and second paragraph of Dumais et al. As shown, the multiplicative scoring function P(L1)*P(L2) is used to combine results from a first tag level and a second level, with a requirement for classification being that the **product** of a first prediction probability corresponding with a first tag at the first tag level and a second prediction probability corresponding with a second tag at the second tag level meets or exceeds a single threshold p=0.2. As long as P(L1)*P(L2) > p = 0.2, even if the matches individually exceed only one level, then the item can be classified as having both the first tag and the second tag. The Boolean scoring function

P(L1)&&P(L2) is used to combine results from a first tag level and the second tag level, with a requirement for classification being that the first prediction probability corresponding with a first tag at the first tag level must meet or exceed a first threshold $p_1 = 0.2$, **and** that the second prediction probability corresponding with a second tag at the second tag level must meet or exceed a second threshold $p_2 = 0.5$.

The Applicants respectfully submit the present claim amendments distinguish from the cited references. Specifically, independent claim 1 as present amended recites the claim features of "**add all parent tags associated with a tag of the plurality of tags to the prediction list** based on a comparison between the prediction probability value for the tag and a first predetermined threshold value". The step of adding **all** parent tags according to the claim feature outlined above aims to solve the problem outlined in [0044] of the present application, which states "Ideally, the classifier should predict the hierarchical representation of tags, for example, for a predicted child tag every parent tag should also be predicted. However it is observed that for some tags returned by the classifier function, the corresponding parent tags [are] not returned. This happens when either the tag is wrongly predicted (false positive) or when some of the correct tags (false negatives) are not predicted." In particular, as discussed in Algorithm 1 and [0048]-[0054], for a predicted child tag, if a prediction probability for the child tag exceeds a first threshold (α, which in [0082] was found to return favorable results when set at α = 0.9), then the system adds **all** parent tags associated with the child tag are added to the prediction list.

In contrast, Dumais does not disclose, teach or suggest adding **all** parent tags associated with a tag of the plurality of tags to the prediction list based on a comparison between the prediction probability value for the tag and a first predetermined threshold value. When utilizing the multiplicative approach P(L1)*P(L2) >p, Dumais evaluates inclusion of a parent tag or a child tag on a **case-by-case basis** by comparing the product of the first prediction probability and the second prediction probability with a threshold value and adding both tags if the product meets or exceeds the threshold value. However, the system of Dumais compares each parent tag with each child tag, which can be computationally expensive and in some cases, unnecessary. In contrast,

the present system adds ALL parent tags associated with a child tag if the child tag has a high enough probability prediction.

Applicant recognizes that using the system of Dumais, the second prediction probability of the child tag can be high enough to cause the product of the first prediction probability and the second prediction probability to exceed the threshold value even if the first prediction probability associated with the parent tag is low. However, in some cases, the system of Dumais can still fail to properly add parent tags if the prediction probability is too low. An example scenario is outlined below:

Suppose there is a tag hierarchy that includes parent tags class#1 and class#2, and a child tag example#1 that belongs to both parent tags class#1 and class#2. A first discussion post about example#1 is correctly tagged by the classifier as belonging to child tag example#1 with a probability of 0.93. The classifier also tags the first discussion post with class#1 with a 0.9 probability and belonging to class#1 with a 0.05 probability due to the nature of the discussion in the post. According to [0044], the discussion post should ideally be classified under both parent tags class#1 and class#2. Given the multiplicative rule of Dumais:

$$P(L1)*P(L2) = 0.93 \times 0.1 = 0.093$$

which is below the allowable threshold of $p = 0.2$. Thus, Dumais would not add class#2 as the parent tag, which is inconsistent with the problem outlined in [0044] or the solution outlined in lines 2 and 6-8 of Algorithm 1 and [0050] of the present application.

In contrast, with the evaluation in Algorithm 1, the present application would determine:

$$P(L2) = 0.93 > (\alpha = 0.9)$$

which is above the allowable "add parent" threshold of $\alpha = 0.9$. Thus, the present application would add class#2 as the parent tag, which is consistent with the problem outlined in [0044] and the solution outlined in lines 2 and 6-8 of Algorithm 1 and [0050] of the present application.

Further, suppose that a second discussion post is included in the same dataset and the second discussion post about example#1 is correctly tagged by the classifier as belonging to parent tag class#2. The system, having previously recognized that the

child tag example#2 also belongs to parent tag class#2 based on the first discussion post, will add **all** known parent tags of example#1 to the second discussion post including both parent tags class#1 and class#2. In contrast, the system of Dumais will only add a parent tag to a discussion post if the combined product of the probabilities of the parent tag and the child tag meet or exceed the threshold value, even if the parent tag is correctly identified as related to the child tag for another discussion post.

The Applicants respectfully submit that the combination of Dumais, Dalins, Otsuka, Grubb, Hassan, Khanzode, Chawla, Otsuka, Rathod, and Chari do not teach or suggest all of the claimed elements of the present invention.

In sum, claim 1 is presently amended to clarify **wherein the prediction list includes a prediction probability value for each tag of the plurality of tags; and** wherein the system **adds all parent tags associated with a tag of the plurality of tags to the prediction list based on a comparison between the prediction probability value for the tag and a first predetermined threshold value**. Further, dependent claims 3 and 4 have been amended to provide further clarification that **all** parent tags are added if the child tag exceeds a first probability threshold, and that **all** child tags are removed if the parent tag falls below a second probability threshold. Claim 2 has been amended for antecedent basis. Independent claim 18 has been amended similarly to independent claim 1 and claims 19 and 20 have been amended for antecedent basis. The features of independent claims 1 and 18 including the present clarifying claim amendments are respectfully absent from Dumais, Dalins, Otsuka, Grubb, Hassan, Khanzode, Chawla, Otsuka, Rathod, and Chari and are non-obvious in view of these references. Claims 12-17 have been cancelled.

Accordingly, in view of the foregoing, the Applicants respectfully submit the system as recited in independent claims 1 and 18 as presently amended is not taught or suggested by Dumais, Dalins, Otsuka, Grubb, Hassan, Khanzode, Chawla, Otsuka, Rathod, and Chari, the claims are believed to be non-obvious, and withdrawal of the 103 rejections is respectfully requested.

CONCLUSION

By the present response, the Applicants respectfully submit that the claims are in condition for allowance, and notification to that effect is earnestly requested. The Examiner is invited to telephone Applicants' representative at (602) 650-2096 to facilitate prosecution of this application. If necessary, please charge any additional fees or credit overpayment to Deposit Account No. 50-1662.

Respectfully submitted,

POLSINELLI PC

Date: 13 August 2021

By: /Timothy D. Fontes

Ari M. Bai, Reg. No.: 70,801
One E. Washington Street, Suite 1200
Phoenix, AZ 85004
Direct Phone: (602) 650-2096
E-Mail: uspt@polsinelli.com
Attorney for Applicants

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

NOTICE OF ALLOWANCE AND FEE(S) DUE

27148 7590 11/12/2021

POLSINELLI PC
PO Box 140310
KANSAS CITY, MO 64114-0310

EXAMINER
THOMAS-HOMESCU, ANNE L

ART UNIT	PAPER NUMBER
2656	

DATE MAILED: 11/12/2021

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/405,612	05/07/2019	Revanth Patil	055743-621457 (M18-191P)	4562

TITLE OF INVENTION: SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB

APPLN. TYPE	ENTITY STATUS	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	MICRO	$300	$0.00	$0.00	$300	02/14/2022

THE APPLICATION IDENTIFIED ABOVE HAS BEEN EXAMINED AND IS ALLOWED FOR ISSUANCE AS A PATENT. PROSECUTION ON THE MERITS IS CLOSED. THIS NOTICE OF ALLOWANCE IS NOT A GRANT OF PATENT RIGHTS. THIS APPLICATION IS SUBJECT TO WITHDRAWAL FROM ISSUE AT THE INITIATIVE OF THE OFFICE OR UPON PETITION BY THE APPLICANT. SEE 37 CFR 1.313 AND MPEP 1308.

THE ISSUE FEE AND PUBLICATION FEE (IF REQUIRED) MUST BE PAID WITHIN THREE MONTHS FROM THE MAILING DATE OF THIS NOTICE OR THIS APPLICATION SHALL BE REGARDED AS ABANDONED. THIS STATUTORY PERIOD CANNOT BE EXTENDED. SEE 35 U.S.C. 151. THE ISSUE FEE DUE INDICATED ABOVE DOES NOT REFLECT A CREDIT FOR ANY PREVIOUSLY PAID ISSUE FEE IN THIS APPLICATION. IF AN ISSUE FEE HAS PREVIOUSLY BEEN PAID IN THIS APPLICATION (AS SHOWN ABOVE), THE RETURN OF PART B OF THIS FORM WILL BE CONSIDERED A REQUEST TO REAPPLY THE PREVIOUSLY PAID ISSUE FEE TOWARD THE ISSUE FEE NOW DUE.

HOW TO REPLY TO THIS NOTICE:

I. Review the ENTITY STATUS shown above. If the ENTITY STATUS is shown as SMALL or MICRO, verify whether entitlement to that entity status still applies.

If the ENTITY STATUS is the same as shown above, pay the TOTAL FEE(S) DUE shown above.

If the ENTITY STATUS is changed from that shown above, on PART B - FEE(S) TRANSMITTAL, complete section number 5 titled "Change in Entity Status (from status indicated above)".

For purposes of this notice, small entity fees are 1/2 the amount of undiscounted fees, and micro entity fees are 1/2 the amount of small entity fees.

II. PART B - FEE(S) TRANSMITTAL, or its equivalent, must be completed and returned to the United States Patent and Trademark Office (USPTO) with your ISSUE FEE and PUBLICATION FEE (if required). If you are charging the fee(s) to your deposit account, section "4b" of Part B - Fee(s) Transmittal should be completed and an extra copy of the form should be submitted. If an equivalent of Part B is filed, a request to reapply a previously paid issue fee must be clearly made, and delays in processing may occur due to the difficulty in recognizing the paper as an equivalent of Part B.

III. All communications regarding this application must give the application number. Please direct all communications prior to issuance to Mail Stop ISSUE FEE unless advised to the contrary.

IMPORTANT REMINDER: Maintenance fees are due in utility patents issuing on applications filed on or after Dec. 12, 1980. It is patentee's responsibility to ensure timely payment of maintenance fees when due. More information is available at www.uspto.gov/PatentMaintenanceFees.

PART B - FEE(S) TRANSMITTAL

Complete and send this form, together with applicable fee(s), by mail or fax, or via EFS-Web.

By mail, send to: Mail Stop ISSUE FEE
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450

By fax, send to: (571)-273-2885

INSTRUCTIONS: This form should be used for transmitting the ISSUE FEE and PUBLICATION FEE (if required). Blocks 1 through 5 should be completed where appropriate. All further correspondence including the Patent, advance orders and notification of maintenance fees will be mailed to the current correspondence address as indicated unless corrected below or directed otherwise in Block 1, by (a) specifying a new correspondence address; and/or (b) indicating a separate "FEE ADDRESS" for maintenance fee notifications.

CURRENT CORRESPONDENCE ADDRESS (Note: Use Block 1 for any change of address)

27148 7590 11/12/2021

POLSINELLI PC
PO Box 140310
KANSAS CITY, MO 64114-0310

Note: A certificate of mailing can only be used for domestic mailings of the Fee(s) Transmittal. This certificate cannot be used for any other accompanying papers. Each additional paper, such as an assignment or formal drawing, must have its own certificate of mailing or transmission.

Certificate of Mailing or Transmission

I hereby certify that this Fee(s) Transmittal is being deposited with the United States Postal Service with sufficient postage for first class mail in an envelope addressed to the Mail Stop ISSUE FEE address above, or being transmitted to the USPTO via EFS-Web or by facsimile to (571) 273-2885, on the date below.

(Typed or printed name)
(Signature)
(Date)

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/405,612	05/07/2019	Revanth Patil	055743-621457 (M18-191P)	4562

TITLE OF INVENTION: SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB

APPLN. TYPE	ENTITY STATUS	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	MICRO	$300	$0.00	$0.00	$300	02/14/2022

EXAMINER	ART UNIT	CLASS-SUBCLASS
THOMAS-HOMESCU, ANNE L	2656	704-009000

1. Change of correspondence address or indication of "Fee Address" (37 CFR 1.363).

❑ Change of correspondence address (or Change of Correspondence Address form PTO/SB/122) attached.

❑ "Fee Address" indication (or "Fee Address" Indication form PTO/SB/47; Rev 03-09 or more recent) attached. **Use of a Customer Number is required.**

2. For printing on the patent front page, list
(1) The names of up to 3 registered patent attorneys or agents OR, alternatively,
(2) The name of a single firm (having as a member a registered attorney or agent) and the names of up to 2 registered patent attorneys or agents. If no name is listed, no name will be printed.

1 _____
2 _____
3 _____

3. ASSIGNEE NAME AND RESIDENCE DATA TO BE PRINTED ON THE PATENT (print or type)

PLEASE NOTE: Unless an assignee is identified below, no assignee data will appear on the patent. If an assignee is identified below, the document must have been previously recorded, or filed for recordation, as set forth in 37 CFR 3.11 and 37 CFR 3.81(a). Completion of this form is NOT a substitute for filing an assignment.

(A) NAME OF ASSIGNEE (B) RESIDENCE: (CITY and STATE OR COUNTRY)

Please check the appropriate assignee category or categories (will not be printed on the patent) : ❑ Individual ❑ Corporation or other private group entity ❑ Government

4a. Fees submitted: ❑ Issue Fee ❑ Publication Fee (if required) ❑ Advance Order - # of Copies _____

4b. Method of Payment: *(Please first reapply any previously paid fee shown above)*

❑ Electronic Payment via EFS-Web ❑ Enclosed check ❑ Non-electronic payment by credit card (Attach form PTO-2038)

❑ The Director is hereby authorized to charge the required fee(s), any deficiency, or credit any overpayment to Deposit Account No. _____

5. **Change in Entity Status** (from status indicated above)

❑ Applicant certifying micro entity status. See 37 CFR 1.29

❑ Applicant asserting small entity status. See 37 CFR 1.27

❑ Applicant changing to regular undiscounted fee status.

NOTE: Absent a valid certification of Micro Entity Status (see forms PTO/SB/15A and 15B), issue fee payment in the micro entity amount will not be accepted at the risk of application abandonment.
NOTE: If the application was previously under micro entity status, checking this box will be taken to be a notification of loss of entitlement to micro entity status.
NOTE: Checking this box will be taken to be a notification of loss of entitlement to small or micro entity status, as applicable.

NOTE: This form must be signed in accordance with 37 CFR 1.31 and 1.33. See 37 CFR 1.4 for signature requirements and certifications.

Authorized Signature _____ Date _____

Typed or printed name _____ Registration No. _____

IPR2025-01398
CrowdStrike EX1002 Page 296

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/405,612	05/07/2019	Revanth Patil	055743-621457 (M18-191P)	4562

27148 7590 11/12/2021

POLSINELLI PC
PO Box 140310
KANSAS CITY, MO 64114-0310

EXAMINER	
THOMAS-HOMESCU, ANNE L	
ART UNIT	PAPER NUMBER
2656	

DATE MAILED: 11/12/2021

Determination of Patent Term Adjustment under 35 U.S.C. 154 (b)
(Applications filed on or after May 29, 2000)

The Office has discontinued providing a Patent Term Adjustment (PTA) calculation with the Notice of Allowance.

Section 1(h)(2) of the AIA Technical Corrections Act amended 35 U.S.C. 154(b)(3)(B)(i) to eliminate the requirement that the Office provide a patent term adjustment determination with the notice of allowance. See Revisions to Patent Term Adjustment, 78 Fed. Reg. 19416, 19417 (Apr. 1, 2013). Therefore, the Office is no longer providing an initial patent term adjustment determination with the notice of allowance. The Office will continue to provide a patent term adjustment determination with the Issue Notification Letter that is mailed to applicant approximately three weeks prior to the issue date of the patent, and will include the patent term adjustment on the patent. Any request for reconsideration of the patent term adjustment determination (or reinstatement of patent term adjustment) should follow the process outlined in 37 CFR 1.705.

Any questions regarding the Patent Term Extension or Adjustment determination should be directed to the Office of Patent Legal Administration at (571)-272-7702. Questions relating to issue and publication fee payments should be directed to the Customer Service Center of the Office of Patent Publication at 1-(888)-786-0101 or (571)-272-4200.

OMB Clearance and PRA Burden Statement for PTOL-85 Part B

The Paperwork Reduction Act (PRA) of 1995 requires Federal agencies to obtain Office of Management and Budget approval before requesting most types of information from the public. When OMB approves an agency request to collect information from the public, OMB (i) provides a valid OMB Control Number and expiration date for the agency to display on the instrument that will be used to collect the information and (ii) requires the agency to inform the public about the OMB Control Number's legal significance in accordance with 5 CFR 1320.5(b).

The information collected by PTOL-85 Part B is required by 37 CFR 1.311. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 30 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, Virginia 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, Virginia 22313-1450. Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

IPR2025-01398
CrowdStrike EX1002 Page 298

Notice of Allowability	Application No. 16/405,612	Applicant(s) Patil et al.	
	Examiner ANNE L THOMAS-HOMESCU	Art Unit 2656	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address--

All claims being allowable, PROSECUTION ON THE MERITS IS (OR REMAINS) CLOSED in this application. If not included herewith (or previously mailed), a Notice of Allowance (PTOL-85) or other appropriate communication will be mailed in due course. **THIS NOTICE OF ALLOWABILITY IS NOT A GRANT OF PATENT RIGHTS.** This application is subject to withdrawal from issue at the initiative of the Office or upon petition by the applicant. See 37 CFR 1.313 and MPEP 1308.

1. ☑ This communication is responsive to 13 August 2021.
   ☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2. ☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

3. ☑ The allowed claim(s) is/are 1-11 and 18-20 . As a result of the allowed claim(s), you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information , please see **http://www.uspto.gov/patents/init_events/pph/index.jsp** or send an inquiry to **PPHfeedback@uspto.gov.**

4. ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

   Certified copies:

   a) ☐ All b) ☐ Some* c) ☐ None of the:

   1. ☐ Certified copies of the priority documents have been received.
   2. ☐ Certified copies of the priority documents have been received in Application No. _____ .
   3. ☐ Copies of the certified copies of the priority documents have been received in this national stage application from the International Bureau (PCT Rule 17.2(a)).

   * Certified copies not received: _____ .

Applicant has THREE MONTHS FROM THE "MAILING DATE" of this communication to file a reply complying with the requirements noted below. Failure to timely comply will result in ABANDONMENT of this application.
THIS THREE-MONTH PERIOD IS NOT EXTENDABLE.

5. ☐ CORRECTED DRAWINGS (as "replacement sheets") must be submitted.
   ☐ including changes required by the attached Examiner's Amendment / Comment or in the Office action of Paper No./Mail Date _____ .

   Identifying indicia such as the application number (see 37 CFR 1.84(c)) should be written on the drawings in the front (not the back) of each sheet. Replacement sheet(s) should be labeled as such in the header according to 37 CFR 1.121(d).

6. ☐ DEPOSIT OF and/or INFORMATION about the deposit of BIOLOGICAL MATERIAL must be submitted. Note the attached Examiner's comment regarding REQUIREMENT FOR THE DEPOSIT OF BIOLOGICAL MATERIAL.

Attachment(s)

1. ☐ Notice of References Cited (PTO-892)
2. ☐ Information Disclosure Statements (PTO/SB/08), Paper No./Mail Date _____.
3. ☐ Examiner's Comment Regarding Requirement for Deposit of Biological Material _____.
4. ☐ Interview Summary (PTO-413), Paper No./Mail Date. _____.
5. ☐ Examiner's Amendment/Comment
6. ☑ Examiner's Statement of Reasons for Allowance
7. ☐ Other _____.

/ANNE L THOMAS-HOMESCU/
Primary Examiner, Art Unit 2656

DETAILED ACTION

1. This communication is in response to the claims filed on 13 August 2021.

2. The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.

3. Claims 1-11 and 18-20 are pending and have been examined, and the Examiner determines that this action is in condition for Allowance.

Reasons for Allowance

1. The following is a statement of reasons for the indication of allowable subject matter:

The closest prior art of record includes over "Hierarchical Classification of Web Content" (Dumais et al.), "Criminal motivation on the dark web: A categorization model for law enforcement" (Dalins et al.), and further in view of US 20080189316, hereinafter referred to as Otsuka. Dumais et al. discloses a computer-implemented system for improving classification arising from cyber environments, comprising: access a ground truth dataset generated from web forum information, the ground truth dataset defining a predetermined tag hierarchy including predetermined tag labels corresponding topics of the web forum information (Dumais et al., p. 258, Web Data Set and Application section. Dumais et al., p. 259, Text Classification Using SVMs section, 1st para.); access data related to a discussion topic for classification from a web forum (Dumais et al., p. 259, Text Classification Using SVMs section, 1st para.); extract a set of features from the data by assigning word vectors or paragraph vectors to the data (See Dumais et al., fig.

1. Dumais et al., p.259, right col., 2nd para.); and apply a machine classifier to the set of features to generate a prediction list of tags for classifying the discussion topic (Dumais et al., p. 259, Text Classification Using SVMs section, 2nd para.), wherein the machine classifier is trained to maintain the predetermined tag hierarchy by regulating modification to parent and child tags associated with a predicted tag of the prediction list of tags based on a prediction probability of the predicted tag and predetermined tag thresholds assigned to the parent and child tags (Dumais et al., p. 260, SVM Parameters section, 2nd para.).

Dumais et al., though, does not disclose classification of criminal activities arising from dark/deep web environments or a processor. Dalins et al. is cited to disclose classification of criminal activities arising from dark/deep web environments (Dalins et al., p. 68, left col. See also Dalins et al., fig. 4 and 5.).

However, none of the prior art, either alone or in combination, teaches:

"apply a machine classifier to the set of features to generate a prediction list of tags for classifying the discussion topic, wherein the prediction list includes a prediction probability value for each tag of the plurality of tags; and

add all parent tags associated with a tag of the plurality of tags to the prediction list based on a comparison between the prediction probability value for the tag and a first predetermined threshold value."

Conclusion

1. The prior art made of record and not relied upon is considered pertinent to the Applicant's disclosure. See PTO 892 from previous office action(s).

Any inquiry concerning this communication or earlier communications from the examiner should be directed to ANNE L THOMAS-HOMESCU whose telephone number is (571)272-0899. The examiner can normally be reached on Mon-Fri 8-6.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Bhavesh Mehta can be reached on 5712727453. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/ANNE L THOMAS-HOMESCU/
Primary Examiner, Art Unit 2659

Issue Classification	Application/Control No.	Applicant(s)/Patent Under Reexamination
	16/405,612	Patil et al.
	Examiner	**Art Unit**
	ANNE L THOMAS-HOMESCU	2656

CPC						
Symbol					**Type**	**Version**
G06F	/	40	/	30	F	2020-01-01
G06N	/	20	/	00	I	2019-01-01
G06F	/	16	/	951	I	2019-01-01
G06F	/	40	/	221	I	2020-01-01
G06F	/	40	/	284	I	2020-01-01

CPC Combination Sets							
Symbol					**Type**	**Set**	**Ranking**	**Version**
	/		/					

NONE		Total Claims Allowed:	
(Assistant Examiner)	(Date)	14	
/ANNE L THOMAS-HOMESCU/ Primary Examiner, Art Unit 2656	06 November 2021	O.G. Print Claim(s)	O.G. Print Figure
(Primary Examiner)	(Date)	1	10

U.S. Patent and Trademark Office

Part of Paper No.: 20211106

IPR2025-01398
CrowdStrike EX1002 Page 303

Issue Classification	Application/Control No.	Applicant(s)/Patent Under Reexamination
	16/405,612	Patil et al.
	Examiner	**Art Unit**
	ANNE L THOMAS-HOMESCU	2656

INTERNATIONAL CLASSIFICATION				
CLAIMED				
G06F40/30	/	40	/	30
G06N20/00	/	20	/	00
G06F16/951	/	16	/	951
G06F40/221	/	40	/	221
G06F40/284	/	40	/	284

NON-CLAIMED				
	/		/	

US ORIGINAL CLASSIFICATION	
CLASS	**SUBCLASS**

CROSS REFERENCES(S)						
CLASS	**SUBCLASS (ONE SUBCLASS PER BLOCK)**					

NONE		Total Claims Allowed:	
(Assistant Examiner)	(Date)	14	
/ANNE L THOMAS-HOMESCU/ Primary Examiner, Art Unit 2656	06 November 2021	O.G. Print Claim(s)	O.G. Print Figure
(Primary Examiner)	(Date)	1	10

Issue Classification	Application/Control No.	Applicant(s)/Patent Under Reexamination
	16/405,612	Patil et al.
	Examiner	Art Unit
	ANNE L THOMAS-HOMESCU	2656

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIMS

Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original
1	1	10	10	13	19										
2	2	11	11	14	20										
3	3		12												
4	4		13												
5	5		14												
6	6		15												
7	7		16												
8	8		17												
9	9	12	18												

NONE		Total Claims Allowed:	
(Assistant Examiner)	(Date)	14	
/ANNE L THOMAS-HOMESCU/ Primary Examiner, Art Unit 2656	06 November 2021	O.G. Print Claim(s)	O.G. Print Figure
(Primary Examiner)	(Date)	1	10

U.S. Patent and Trademark Office

Part of Paper No.: 20211106

IPR2025-01398
CrowdStrike EX1002 Page 305

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	16/405,612	Patil et al.
	Examiner	Art Unit
	ANNE L THOMAS-HOMESCU	2656

CPC - Searched*		
Symbol	Date	Examiner
G06F40/30,221,284 16/951; G06N20/00,10,20, 5/003, 3/084	02/19/2021	ATH
G06F40/30,221,284 16/951; G06N20/00,10,20, 5/003, 3/084	11/06/2021	ATH

CPC Combination Sets - Searched*		
Symbol	Date	Examiner

US Classification - Searched*			
Class	Subclass	Date	Examiner

* See search history printout included with this form or the SEARCH NOTES box below to determine the scope of the search.

Search Notes		
Search Notes	Date	Examiner
See attached EAST and Google Scholar searches.	02/19/2021	ATH
See attached EAST and Google Scholar searches.	11/06/2021	ATH

Interference Search			
US Class/CPC Symbol	US Subclass/CPC Group	Date	Examiner
	G06F40/30,221,284 16/951; G06N20/00,10,20, 5/003, 3/084	11/06/2021	ATH

/ANNE L THOMAS-HOMESCU/ Primary Examiner, Art Unit 2656	

IPR2025-01398
CrowdStrike EX1002 Page 306

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
	16/405,612	Patil et al.
	Examiner	Art Unit
	ANNE L THOMAS-HOMESCU	2656

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

CLAIMS

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	02/17/2021	11/06/2021							
	1	✓	=							
	2	✓	=							
	3	✓	=							
	4	✓	=							
	5	✓	=							
	6	✓	=							
	7	✓	=							
	8	✓	=							
	9	✓	=							
	10	✓	=							
	11	✓	=							
	12	✓	-							
	13	✓	-							
	14	✓	-							
	15	✓	-							
	16	✓	-							
	17	✓	-							
	18	✓	=							
	19	✓	=							
	20	✓	=							

EAST Search History

EAST Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
S60	4	("10176438").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:33
S61	7	("10313385").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:36
S62	2	("10437945").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:37
S63	2	("20190349393").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:38
S64	3	("20200036743").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:39
S65	5	("2019089389").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:40

S66	4	("2019157335").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:41
S67	2,734	(deep near3 web)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:47
S68	3,521	((deep dark) near3 web)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:48
S69	4,198	((deep dark) near3 (web internet))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:51
S70	8,632	((deep dark) near7 (web internet))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:51
S71	2,196	((deep dark) near3 (web internet)) and (forum$1 discuss$4 blog$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:53
S72	34	((deep dark) near3 (web internet)) and (forum$1 discuss$4 blog$4 meeting$1 conference$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:56

S73	34	((deep dark) near3 (web internet)) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:57
S74	34	((deep dark) near3 (web internet) D2web) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 09:59
S75	50	((deep dark) near3 (web internet) D2web) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 10:02
S76	11,294	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 10:05
S77	6	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (parent child hierarch$3) same (train$3 learn$3 adapt$4 adjust$4) and (criminal$3 illegal$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 10:57
S78	32	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (parent child hierarch$3) same (train$3 learn$3 adapt$4 adjust$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 10:58
S79	6	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media chat$4) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (parent child hierarch$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 10:59

		same (train$3 learn$3 adapt$4 adjust$4) and (criminal$3 illegal$4)				
S80	5,248	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) and (index$4 (dark near3(web internet cyber)))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:08
S81	886	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) and (index$4 (dark near3(web internet cyber)))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:11
S82	31	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child) and (index$4 (dark near3(web internet cyber)))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:29
S83	107	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) and (index$4 (dark near3 (web internet cyber)))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:30
S84	76	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) and (index$4 (dark near3 (web internet cyber))) NOT S82	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:45

S85	2	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) and (deep near3 web) NOT S82	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:53
S86	2	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child) and (deep near3 web)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:56
S87	144	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:56
S88	50	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:58
S89	37	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 12:59

S90	50	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 13:05
S91	52	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media post$3) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 13:12
S92	238	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media post$3) same (topic$1 subject$1 title$1 theme$1 classification classif$4 class$2 categor$3 categori$1ation group$4) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 13:15
S93	1,797	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media post$3) same (topic$1 subject$1 title$1 theme$1 classification classif$4 class$2 categor$3 categori$1ation group$4) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 13:16
S94	110	(dark$1net ((dark deep) near3 web)) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO;	OR	OFF	2021/02/16 13:34

			DERWENT; IBM_TDB			
S95	1,005	(dark$1net ((dark deep) near3 (web internet cyber network$1))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 13:55
S96	895	(dark$1net ((dark deep) near3 (web internet cyber network$1))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged) NOT S94	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 13:58
S97	31	(dark$1net ((dark deep) near3 (web internet cyber network$1))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) NOT S94	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 14:02
S98	0	(dark$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) NOT S94	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 14:05
S99	53	(dark$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged identifier$1 identification$1 identif$4) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1) NOT S94	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 14:08
S100	55	(dark$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged identifier$1 identification$1 identif$4 cluster$3 group$4) same (forum$1 discuss$4 blog$4 meeting$1	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 14:24

		conference$1 panel$4 media page$1 site$1 web$1site$1) NOT S94				
S101	46	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 16:48
S102	166	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (hierarch$3 parent child subset$4 tree leaf leaves) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 16:57
S103	1,092	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 17:18
S104	36	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 17:22

		confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec) and (ground near3 truth)				
S105	941	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec) and ((ground near3 truth) training)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 17:52
S106	7	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (hierarch$3 sub$1categor$3 sub$1class$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec) and (ground near3 truth)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 18:03
S107	48	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (hierarch$3 sub$1categor$3 sub$1class$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec) and ((ground near3 truth) training)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 18:04
S108	52	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1	US-PGPUB; USPAT;	OR	OFF	2021/02/16 18:07

		conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (hierarch$3 sub$1categor$3 sub$1class$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec)	USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB			
S109	77	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (hierarch$3 sub$1categor$3 sub$1class$3 parent child) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) and (vector$1 word2vec)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 18:10
S110	2	("7349917").PN.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/16 18:25
S111	110	(dark$1net ((dark deep) near3 web)) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 13:22
S112	51	(dark$1net crime$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged group$3) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 13:22
S113	0	(dark$1net crime$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3)	US-PGPUB; USPAT; USOCR;	OR	OFF	2021/02/17 13:25

		same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged group$3) same (title$1) and (process$3) NOT S111	FPRS; EPO; JPO; DERWENT; IBM_TDB			
S114	51	(dark$1net crime$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged group$3) and (process$3) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 13:26
S115	10	(dark$1net crime$1net ((dark deep) near3 (web internet cyber))) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 13:32
S116	48	(dark$1net crime$1net ((dark deep) near3 (web internet cyber))) and (crime$1 criminal$3 illegal$3 fraud faudulent$3 theft$1) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 13:33
S117	0	(dark$1net crime$1net ((dark deep) near3 (web internet cyber))) and (crime$1 criminal$3 illegal$3 fraud faudulent$3 theft$1) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 title$1) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 14:37
S118	63,544	(classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 title$1) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 14:44
S119	3,622	(classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 subject$1) same (title$1) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 14:45
S120	1,690	(web$1site$1 webpage$1) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged	US-PGPUB; USPAT; USOCR;	OR	OFF	2021/02/17 15:25

		cluster$3) near7 (topic$1 subject$1) same (title$1) NOT S111	FPRS; EPO; JPO; DERWENT; IBM_TDB			
S121	575	(web$1site$1 webpage$1) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 subject$1) near7 (title$1) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 15:25
S122	40	(web$1site$1 webpage$1) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 subject$1) near7 (title$1) same (vector$1 feature$1 word2vec) NOT S111	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 15:26
S123	0	(web internet cyber$4) and (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media page$1 site$1 web$1site$1 web$1page$1) same (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 16:27
S124	4	(web internet cyber$4) and (topic$1 subject$1 title$1 theme$1) near7 (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/17 16:28
S125	435	(web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 08:47

S126	1	(web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 08:50
S127	1	(web internet cyber$4) and (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 08:51
S128	1	(probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 08:51
S129	1	(web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (probabilit$3 confidence rank$3 best certaint$3 threshold$3 scor$3 metric$4) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 09:13
S130	1,030	(web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same (character$1 word$1) near7 (remov$3 delet$3 ignor$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:19
S131	855	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same (character$1 word$1) near7 (remov$3 delet$3 ignor$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:25

S132	118	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same (character$1) same (word$1) near7 (remov$3 delet$3 ignor$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:26
S133	315	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same (character$1) same (word$1) same (remov$3 delet$3 ignor$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:27
S134	50	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same ((character$1) same (word$1)) near7 (remov$3 delet$3 ignor$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:31
S135	1	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same ((character$1) same (word$1)) near7 (remov$3 delet$3 ignor$4) and (root$1) same (stem$4) same (lemmati$1ation lemma$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:35
S136	1	(G10L.cpcl. G06F.cpcl.) and (pars$4 segmentation segment$3) same (tokeni$1ation token$4 label$4 tag$4) same ((character$1) same (word$1)) near7 (remov$3 delet$3 ignor$4) and (root$1) same (stem$4) same (lemmati$1ation lemma$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 10:37
S137	131	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2	US-PGPUB; USPAT; USOCR;	OR	OFF	2021/02/18 10:38

		classification categori$1ation categori$4 group$3 cluster$3) and (root$1) same (stem$4) same (lemmati$1ation lemma$1)	FPRS; EPO; JPO; DERWENT; IBM_TDB			
S138	28,160	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (crawl$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:48
S139	6,829	(G10L.cpcl. G06F.cpcl.) and (web internet cyber$4) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (crawl$3) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:49
S140	14,864	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (crawl$3) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media site$1 web$1site$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:49
S141	5,259	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (crawl$3) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (internet web site$1 web$1site$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:51
S142	2,342	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) and (crawler) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (internet web site$1 web$1site$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:53
S143	838	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (crawler) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (internet web site$1 web$1site$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:54

S144	354	(G10L.cpcl. G06F.cpcl.) and (classif$4 class$2 classification categori$1ation categori$4) same (crawler) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (internet web site$1 web$1site$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 11:57
S145	56	(G10L.cpcl. G06F.cpcl.) and (classif$4 class$2 classification categori$1ation categori$4) same (crawl$3) same (spider$4) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (internet web site$1 web$1site$1 html)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 12:25
S146	110	(dark$1net ((dark deep) near3 web)) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 13:29
S147	0	(web$1site$1 webpage$1) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 subject$1) near7 (title$1) same (date$1) same (reputation$1 review$4 comment$4) same (author$1 writer$1) same (vector$1 feature$1 word2vec) NOT S146	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 13:30
S148	7	(web$1site$1 webpage$1) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) near7 (topic$1 subject$1) near7 (title$1) same (date$1) same (reputation$1 review$4 comment$4) same (author$1 writer$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 13:32
S149	20	(web$1site$1 webpage$1) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged cluster$3) same (topic$1 subject$1) near7 (title$1) same (date$1) same (reputation$1 review$4 comment$4) same (author$1 writer$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 13:36
S150	3	(G10L.cpcl. G06F.cpcl.) and (classif$4 class$2 classification categori$1ation categori$4) same (crawl$3) same (spider$4) same (forum$1 discuss$4 blog$4 meeting$1 conference$1 panel$4 media) same (internet web	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO;	OR	OFF	2021/02/18 13:56

		site$1 web$1site$1 html) same (train$3 (ground near3 truth))	DERWENT; IBM_TDB			
S151	1,784	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 16:23
S152	1,785	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 16:24
S153	273	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (semi$1supervis$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 16:26
S154	1	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (semi$1supervis$4) same (synthetic$4 synthes$4 artificial$4) near7 (sampl$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 17:57
S155	31	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (nearest near3 neighbor$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 18:50
S156	1	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (nearest near3 neighbor$1) same (elastic$4)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 21:08
S158	4	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation	US-PGPUB; USPAT;	OR	OFF	2021/02/18 21:13

		categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4)	USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB			
S159	4	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (train$3 adaptation adapt$3 adjust$4 learn$3 updat$3) same ((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4 BM25)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 21:33
S160	25	((ground near3 truth) training) same (imbalance$1 minorit$3) same (elastic$4 BM25)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 21:35
S161	3,029	(imbalance$1 minorit$3) same (elastic$4 BM25)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 21:44
S162	303	(label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (imbalance$1 minorit$3) same (elastic$4 BM25)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 21:44
S163	3	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (imbalance$1 minorit$3) same (elastic$4 BM25)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/18 21:45
S164	108	(dark$1net ((dark deep) near3 web)) and (criminal$3 illegal$3 fraud faudulent$3) same (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged) and (process$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/19 09:07
S165	1,898	(dark$1net ((dark deep) near3 web)) and (classification classif$4	US-PGPUB; USPAT; USOCR;	OR	OFF	2021/02/19 09:20

		categori$1ation categor$3 label$4 tag tagging tagged) and (process$3)	FPRS; EPO; JPO; DERWENT; IBM_TDB			
S166	287	(dark$1net ((dark deep) near3 web)) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged) near7 (updat$3 adjust$4 adaptation adapt$3) and (process$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/19 09:59
S167	292	(dark$1net ((dark deep) near3 web)) and (classification classif$4 categori$1ation categor$3 label$4 tag tagging tagged) near7 (updat$3 adjust$4 adaptation adapt$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/02/19 10:30
S168	9,971	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (list$3 rank$4 table$1) same (probabilit$3 likelihood$1) same (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/11/04 15:42
S169	4,580	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3) same (list$3 rank$4 table$1) same (probabilit$3 likelihood$1) near7 (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/11/04 15:53
S170	1,780	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 title$1) same (list$3 rank$4 table$1) same (probabilit$3 likelihood$1) near7 (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/11/04 15:55
S171	143	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 title$1) near7 (list$3 rank$4 table$1) near7 (probabilit$3 likelihood$1) near7 (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/11/04 15:55
S172	145	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 title$1) near7 (list$3	US-PGPUB; USPAT;	OR	OFF	2021/11/04 16:06

		rank$4 table$1) near7 (probabilit$3 likelihood$1) near7 (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1 range$1 ranging)	USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB			
S173	540	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 title$1 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3 subject$1) near7 (list$3 rank$4 table$1) near7 (probabilit$3 likelihood$1) near7 (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1 range$1 ranging)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/11/04 16:36
S174	395	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 title$1 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3 subject$1) near7 (list$3 rank$4 table$1) near7 (probabilit$3 likelihood$1) near7 (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1 range$1 ranging) NOT S172	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/11/04 17:04
S175	1,201	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 title$1 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3 subject$1) near7 (list$3 rank$4 table$1) near7 (probabilit$3 likelihood$1 weight$4) near7 (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1 range$1 ranging) NOT S172	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/11/04 17:29
S176	16	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 title$1 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3 subject$1) near7 (list$3 rank$4 table$1) near7 (probabilit$3 likelihood$1) near7 (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1 range$1 ranging) same (hierarchic$4 hierarch$3 parent$1 child$4) NOT S172	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/11/04 18:49
S177	30	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 title$1 mark$4 flag$4 subject$1) near7 (list$3 rank$4 table$1) near7 (probabilit$3 likelihood$1) near7 (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1 range$1 ranging) NOT S172	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/11/04 19:11

S178	38	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 title$1 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3 subject$1) near7 (list$3 rank$4 table$1) near7 (probabilit$3 likelihood$1) near7 (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1 range$1 ranging) same (hierarchic$4 hierarch$3 parent$1 child$4 root$1 node$1) NOT S172	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/11/04 19:20
S179	145	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 title$1) near7 (list$3 rank$4 table$1) near7 (probabilit$3 likelihood$1) near7 (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1 range$1 ranging)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/11/04 20:52
S180	105	(G10L.cpcl. G06F.cpcl.) and (label$4 tag$4 mark$4 title$1 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3 subject$1) near7 (list$3 rank$4 table$1) near7 (probabilit$3 likelihood$1) near7 (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1 range$1 ranging) same (hierarchic$4 hierarch$3 parent$1 child$4 root$1 node$1 first) NOT S179	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2021/11/04 20:52

EAST Search History (Interference)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
L1	39	(G10L.cpcl. G06F.cpcl.) and ((label$4 tag$4 mark$4 title$1 mark$4 flag$4 classif$4 class$2 classification categori$1ation categori$4 group$3 cluster$3 subject$1) near7 (list$3 rank$4 table$1) near7 (probabilit$3 likelihood$1) near7 (threshold$1 limit$4 confidence$1 bound$1 value$1 level$1 range$1 ranging) same (hierarchic$4 hierarch$3 parent$1 child$4 root$1 node$1 first)).clm.	US-PGPUB; USPAT	OR	OFF	2021/11/06 10:57

11/6/2021 11:05:22 AM
C:\Users\athomashomescu\Documents\EAST\Workspaces\16405612 - assigning labels to discussion topics shown in online forums on the dark web.wsp

Document Description: Issue Fee Payment (PTO-85B)

Issue Fee Transmittal Form

Application Number	Filing Date	First Named Inventor	Atty. Docket No.	Confirmation No.
16405612	07-May-2019	Revanth Patil	055743-621457 (M18-191P	4562

TITLE OF INVENTION :

SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB

Entity Status	Application Type	Art Unit	Class - Subclass	EXAMINER
Micro	Utility under 35 USC 111(a)	2656	009000	

Issue Fee Due	Publication Due	Total Fee(s) Due	Date Due	Prev. Paid Fee
$300	$0	$300	14-Feb-2022	$0

1.Change of Correspondence Address and/or Indication Of Fee Address (37 CFR 1.33 & 1.363)

Current Correspondence Address:	Current Indicated Fee Address :
27148 POLSINELLI PC PO Box 140310 KANSAS CITY MO 64114-0310 UNITED STATES 816-753-1000 patentdocketing@polsinelli.com	
☐ Change of correspondence address requested, system generated AIA/122-EFS form attached	☐ Fee Address indication requested, system generated SB/47-EFS form attached

2.Entity Status

Change in Entity Status

◉ Applicant certifying micro entity status; system generated Micro Entity certification form attached. See 37 CFR 1.29.
Note: Absent a valid certification of micro entity status, issue fee payment in the micro entity amount will not be accepted at the risk of application abandonment. If this box is checked, you will be prompted to choose a micro entity status on the gross income basis (37 CFR 1.29(a)) or the institution of higher education basis (37 CFR 1.29(d)), and make the applicable certification online.

○ Applicant asserting small entity status. See 37 CFR 1.27.
Note: If the application was previously under micro entity status, checking this box will be taken to be a notification of loss of entitlement to micro entity status.

○ Applicant changing to regular undiscounted fee status.
Note: Checking this box will be taken to be a notification of loss of entitlement to small or micro entity status, as applicable.

IPR2025-01398
CrowdStrike EX1002 Page 329

Document Description: Issue Fee Payment (PTO-85B)

3.The Following Fee(s) Are Submitted:

☒ Issue Fee

☐ I authorize USPTO to apply my previously paid issue fee to the current fees due

☐ Publication Fee

☐ The Director is hereby authorized to apply my previously paid issue fee to the current fee due and to charge deficient fees to Deposit Account Number _____

☐ Advance Order - # of copies _____

☒ If **in addition to** the payment of the issue fee amount submitted with this form, there are any discrepancies in any amount(s) due, the Director is authorized to charge any deficiency or credit any overpayment, to Deposit Account Number 501662 . The **issue fee must be submitted** with this form. **If payment of the issue fee does not accompany this form, checking this box and providing a deposit account number will NOT be effective to satisfy full payment of the fee(s) due.**

4.Firm and/or Attorney Names To Be Printed

NOTE: If no name is listed, no name will be printed
For printing on the patent front page, list to be displayed as entered

1. POLSINELLI PC
2.
3.

5.Assignee Name(s) and Residence Data To Be Printed

PLEASE NOTE: Unless an assignee is identified below, no assignee data will appear on the patent. If an assignee is identified below, the document has been filed for recordation as set forth in 37 CFR 3.11. Completion of this form is NOT a substitute for filing an assignment.

Name	City	State	Country	Category
Arizona Board of Regents on Behalf of Arizona State University	Tempe	ARIZONA	UNITED STATES	corporation

6.Signature

I certify, in accordance with 37 CFR 1.4(d)(4) that I am an attorney or agent registered to practice before the Patent and Trademark Office who has filed and has been granted power of attorney in this application. I also certify that this Fee(s) Transmittal form is being transmitted to the USPTO via EFS-WEB on the date indicated below.

Signature	/Ari M. Bai/	**Date**	02-04-2022
Name	Ari M. Bai	**Registration Number**	38726

Electronic Patent Application Fee Transmittal

Application Number:	16405612
Filing Date:	07-May-2019
Title of Invention:	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB
First Named Inventor/Applicant Name:	Revanth Patil
Filer:	Ari M. Bai/Rebecca Endsley
Attorney Docket Number:	055743-621457 (M18-191P)

Filed as Micro Entity

Filing Fees for Utility under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
UTILITY ISSUE FEE	3501	1	300	300
PUBL. FEE- EARLY, VOLUNTARY, OR NORMAL	1504	1	0	0
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				
Miscellaneous:				
		Total in USD ($)		**300**

Electronic Acknowledgement Receipt

EFS ID:	44910326
Application Number:	16405612
International Application Number:	
Confirmation Number:	4562
Title of Invention:	SYSTEMS AND METHODS FOR AUTOMATICALLY ASSIGNING ONE OR MORE LABELS TO DISCUSSION TOPICS SHOWN IN ONLINE FORUMS ON THE DARK WEB
First Named Inventor/Applicant Name:	Revanth Patil
Customer Number:	27148
Filer:	Ari M. Bai/Rebecca Endsley
Filer Authorized By:	Ari M. Bai
Attorney Docket Number:	055743-621457 (M18-191P)
Receipt Date:	04-FEB-2022
Filing Date:	07-MAY-2019
Time Stamp:	15:20:33
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	DA
Payment was successfully received in RAM	$300
RAM confirmation Number	E202224F20310222
Deposit Account	
Authorized User	

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Issue Fee Payment (PTO-85B)	Web85b.pdf	62785 e15d421ad35fe6b9641d74c15949336a151aa250	no	2
Warnings:					
Information:					
2	Fee Worksheet (SB06)	fee-info.pdf	40371 e3f9bdae2d6529e0eac784fc2d518a4d0d6a409e	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	103156		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	ISSUE DATE	PATENT NO.	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/405,612	03/15/2022	11275900	055743-621457 (M18-191P)	4562

27148 7590 02/23/2022

POLSINELLI PC
PO Box 140310
KANSAS CITY, MO 64114-0310

ISSUE NOTIFICATION

The projected patent number and issue date are specified above.

Determination of Patent Term Adjustment under 35 U.S.C. 154 (b)
(application filed on or after May 29, 2000)

The Patent Term Adjustment is 151 day(s). Any patent to issue from the above-identified application will include an indication of the adjustment on the front page.

If a Continued Prosecution Application (CPA) was filed in the above-identified application, the filing date that determines Patent Term Adjustment is the filing date of the most recent CPA.

Applicant will be able to obtain more detailed information by accessing the Patent Application Information Retrieval (PAIR) WEB site (http://pair.uspto.gov).

Any questions regarding the Patent Term Extension or Adjustment determination should be directed to the Office of Patent Legal Administration at (571)-272-7702. Questions relating to issue and publication fee payments should be directed to the Application Assistance Unit (AAU) of the Office of Data Management (ODM) at (571)-272-4200.

INVENTOR(s) (Please see PAIR WEB site http://pair.uspto.gov for additional inventors):

Revanth Patil, Tempe, AZ;
Paulo Shakarian, Tempe, AZ;
Ashkan Aleali, Tempe, AZ;
Ericsson Marin, Tempe, AZ;

APPLICANT(s) (Please see PAIR WEB site http://pair.uspto.gov for additional applicants):

The United States represents the largest, most dynamic marketplace in the world and is an unparalleled location for business investment, innovation, and commercialization of new technologies. The USA offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to encourage and facilitate business investment. To learn more about why the USA is the best country in the world to develop technology, manufacture products, and grow your business, visit SelectUSA.gov.