Machine Learning with Applications 14 (2023) 100508

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Transforming sentiment analysis in the financial domain with ChatGPT

Georgios Fatouros [a,b,*], John Soldatos [b], Kalliopi Kouroumali [c], Georgios Makridis [a], Dimosthenis Kyriazis [a]

[a] Department of Digital Systems, University of Piraeus, Karaoli and Dimitriou 80, Piraeus, 18534, Greece
[b] Innov-Acts Limited, Kolokotroni 6, Nicosia, 1101, Cyprus
[c] Hellenic Telecommunications Organisation S.A., Kifissias Avenue 99, Athens, 15124, Greece

ARTICLE INFO

MSC:
68T01
68T50
91B28
91B30

Keywords:
ChatGPT
Artificial intelligence
Finance
Sentiment analysis
Risk assessment

ABSTRACT

Financial sentiment analysis plays a crucial role in decoding market trends and guiding strategic trading decisions. Despite the deployment of advanced deep learning techniques and language models to refine sentiment analysis in finance, this study breaks new ground by investigating the potential of large language models, particularly ChatGPT 3.5, in financial sentiment analysis, with a strong emphasis on the foreign exchange market (forex). Employing a zero-shot prompting approach, we examine multiple ChatGPT prompts on a meticulously curated dataset of forex-related news headlines, measuring performance using metrics such as precision, recall, f1-score, and Mean Absolute Error (MAE) of the sentiment class. Additionally, we probe the correlation between predicted sentiment and market returns as an addition evaluation approach. ChatGPT, compared to FinBERT, a well-established sentiment analysis model for financial texts, exhibited approximately 35% enhanced performance in sentiment classification and a 36% higher correlation with market returns. By underlining the significance of prompt engineering, particularly in zero-shot contexts, this study spotlights ChatGPT's potential to substantially boost sentiment analysis in financial applications. By sharing the utilized dataset, our intention is to stimulate further research and advancements in the field of financial services.

1. Introduction

1.1. Background and motivation

The financial services sector has long been an early adopter of technological advancements, continually evolving to meet the demands of a rapidly changing global landscape. From the introduction of ATMs (Automatic Teller Machines) and electronic trading to the rise of financial technologies (Fintech), the financial sector has witnessed a transformative shift, particularly with the integration of artificial intelligence (AI) and machine learning (ML) technologies (Arner, Barberis, & Buckley, 2015). The Fintech market has seen significant growth, demonstrating the impact of these novel technologies in finance (Mordor Intelligence, 2022). Such services include credit scoring and risk assessment, fraud detection and prevention, robo-advisory services as well as chatbots for customer service and support (Fatouros et al., 2023; Kotios, Makridis, Fatouros, & Kyriazis, 2022).

Among the many facets of this transformation, sentiment analysis has emerged as a critical tool for understanding market dynamics and predicting future trends. The ability to analyze sentiments can provide valuable insights into the collective mood of the market, thus enabling more informed and strategic decision-making.

Traditionally, financial sentiment analysis has relied on manually curated lexicons and simple ML algorithms (Schumaker & Chen, 2009). However, with the rapid progress in Natural Language Processing (NLP), more sophisticated techniques are now available. Deep learning-based models, such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin, Chang, Lee, & Toutanova, 2018) and its financial domain-specific counterpart FinBERT (Liu, Huang, Huang, Li, & Zhao, 2021), have significantly improved sentiment analysis accuracy and reliability.

Despite these advancements, the financial domain presents unique and multifaceted challenges to sentiment analysis. Financial texts, particularly news headlines, are often saturated with domain-specific terminology and nuanced sentiments, complicating the task of sentiment classification (Loughran & McDonald, 2011). A distinctive feature of forex news is the intertwining of sentiments about multiple currencies

* Corresponding author at: Department of Digital Systems, University of Piraeus, Karaoli and Dimitriou 80, Piraeus, 18534, Greece.
E-mail addresses: gfatouros@unipi.gr (G. Fatouros), jsoldat@innov-acts.com (J. Soldatos), kkouroumali@ote.gr (K. Kouroumali), gmakridis@unipi.gr (G. Makridis), dimos@unipi.gr (D. Kyriazis).

https://doi.org/10.1016/j.mlwa.2023.100508
Received 30 May 2023; Received in revised form 15 October 2023; Accepted 2 November 2023
Available online 4 November 2023
2666-8270/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

within a single statement. Consider, for instance, the headline "CAD is one of the better places to sell the US dollar on a pullback". Conventional sentiment analysis tools, might predict a positive sentiment for this statement, which is accurate concerning the Canadian dollar but misleading for the forex pair USD/CAD. Such models mainly focus on sentiment classification, often failing to infer the subject of the text, i.e., the financial instruments related to the sentiment, which is a crucial aspect of financial analysis.

In addition to this, conventional sentiment analysis models often lack the ability to adjust their output based on specific use-case context, further limiting their broader applicability (Poria, Cambria, & Gelbukh, 2016). For example, the sentiment expressed in a discussion about a new governmental regulatory policy may differ significantly depending on whether the context is an investor forum discussing potential market impacts or a consumer forum discussing changes in service fees. Typical sentiment analysis models may fail to capture these contextual nuances effectively, offering room for more context-aware models to provide more comprehensive and accurate sentiment analysis in the financial domain (Poria, Cambria, Bajpai, & Hussain, 2017). Thus, addressing these limitations and meeting these specific challenges head-on represents a significant area of opportunity for advancements in financial sentiment analysis.

Among the various AI tools, Generative Pre-trained Transformers (GPT) and their related technologies have demonstrated significant potential in revolutionizing multiple domains, including the financial sector (George & George, 2023). GPT, a state-of-the-art language model developed by OpenAI, is trained on an extensive corpus of data, enabling it to understand and replicate patterns in human language with remarkable precision (Radford, Narasimhan, Salimans, Sutskever, et al., 2018). Since the introduction of the initial GPT model in 2018, OpenAI has consistently improved the architecture by adding new layers, implementing technical adjustments, and extending training on additional datasets, culminating in the latest and most advanced version, GPT-4 (OpenAI, 2023). However, the broader impact of GPT on the public became evident towards the end of 2022 with the release of ChatGPT, a specialized version of the GPT architecture fine-tuned for conversational applications.

ChatGPT offers promising opportunities to improve existing financial applications, such as risk analysis through sentiment analysis. By harnessing the power of ChatGPT's advanced natural language understanding capabilities, financial institutions can significantly enhance the accuracy and depth of sentiment analysis. With its ability to comprehend and interpret complex language patterns, it can more effectively analyze vast amounts of unstructured data, such as news articles and headlines, to provide a comprehensive understanding of market sentiment. This improved sentiment analysis can, in turn, inform decision-making in areas such as investment strategies, risk management, and portfolio optimization, leading to better-informed decisions and potentially higher returns (Chen, De, Hu, & Hwang, 2014; Tetlock, 2007). Additionally, ChatGPT's conversational abilities can be utilized to communicate and explain complex risk analysis findings to both expert and novice users, making financial insights more accessible and actionable (Yue, Au, Au, & Iu, 2023).

These ChatGPT's advanced natural language understanding capabilities carry potential for a wealth of innovation in the financial domain, including financial sentiment analysis. This area, despite its significant implications, remains largely untapped, signaling an open avenue for research and development. The capacity of ChatGPT to interpret complex language patterns, a characteristic often found in unstructured financial data, presents a promising solution for enhancing the depth and accuracy of sentiment analysis in financial contexts.

1.2. Value proposition

The domain of financial sentiment analysis has witnessed the application of various language models (LMs). However, the potential of large language models (LLMs) like ChatGPT 3.5 in the context of the foreign exchange market (forex) remains relatively uncharted, as detailed in Section 2. Forex news, distinct from general financial news, often encapsulates sentiments about two distinct assets within a single statement, introducing an added layer of complexity. Recognizing this intricate nature, our study pioneers the exploration of ChatGPT 3.5's capabilities in discerning such nuanced sentiment cues. To our knowledge, this is the first study that evaluates ChatGPT's capabilities specifically for forex news sentiment analysis.

In our approach, we leverage a zero-shot prompting strategy, which not only assesses ChatGPT's proficiency in interpreting forex-related financial text but also emphasizes its ability to achieve this without domain-specific fine-tuning. This work significantly advances the current understanding of generative AI potential in specialized financial sentiment analysis tasks and their applicability in real-world scenarios.

To ensure a reliable evaluation, we have meticulously curated and annotated a dataset of forex-related news headlines. This dataset, made publicly available, facilitates a comprehensive evaluation of various ChatGPT prompts, thereby highlighting the model's extensive training on diverse datasets and its inherent adaptability in different sectors, including finance. By sharing this annotated dataset, we aim to contribute a valuable resource to the research community, fostering further advancements and innovation in the field of financial services and particularly in financial sentiment analysis.

The evaluation goes beyond traditional metrics, including not only precision, recall, f1-score, and Mean Absolute Error (MAE) of the sentiment class, but also the correlation between predicted sentiment and market returns (Hossin & Sulaiman, 2015). This detailed evaluation extends to sentiment class prediction and sentiment score estimation for both single, multiple news headlines, as well as the complete news article, demonstrating ChatGPT's agility and proficiency in handling intricate financial texts.

Furthermore, our study benchmarks ChatGPT's performance against the state-of-the-art model, FinBERT, and a naive approach, offering a holistic view of its capabilities in the financial domain. By exploring multiple prompts, we emphasize the importance of prompt engineering, especially in zero-shot contexts, to optimize performance and enhance sentiment analysis efficacy.

In essence, as models like ChatGPT become increasingly integrated into real-world applications, our work offers valuable insights that can guide developers and researchers in harnessing the model's potential effectively. The curated dataset and annotations, which we have made publicly available, further solidify our contribution, serving as a useful resource for future research endeavors in this domain.

The remainder of this paper is structured as follows: Section 2 reviews prior work on sentiment analysis in finance and discusses the role of AI tools such as ChatGPT and FinBERT in this domain. Section 3 details our methodology, highlighting our approach to using and evaluating ChatGPT for sentiment classification in the forex market. Section 4 presents our findings, which include performance metrics of sentiment classification and sentiment correlation with market returns. In Section 5, we discuss the implications of our results, the potential applications of ChatGPT in financial services, as well as future research directions. The paper is concluded in Section 6 with a summary of our study's key contributions.

2. Related work

2.1. Sentiment analysis in finance

Sentiment analysis, or opinion mining, refers to the use of computational techniques to extract subjective information, such as opinions and sentiments, from textual data (Bing, 2012). It has been applied to various domains, including customer reviews, social media posts, and news articles. In finance, sentiment analysis is used to gauge market sentiment, an essential aspect that could impact the price

movement of financial instruments (Bollen, Mao, & Zeng, 2011; Siering, Muntermann, & Rajagopalan, 2018).

The idea that sentiment or public mood can influence financial markets dates back to the era of Keynes (1937), who described financial markets as being driven by "animal spirits" or waves of pessimism and optimism. This notion has been substantiated by numerous empirical studies, demonstrating that changes in investor sentiment could significantly affect asset prices and trading volume (Baker & Wurgler, 2007; Tetlock, 2007).

Initially, financial sentiment analysis relied on manually curated financial lexicons, where words were pre-labeled as positive, negative, or neutral (Loughran & McDonald, 2011). However, this approach was often too simplistic, overlooking the context in which words were used. For example, the word "crisis" may have negative connotations in general but could be used in a positive context, such as "the company successfully averted a crisis".

The advent of ML techniques brought significant advancements in sentiment analysis. ML models, trained on labeled financial news articles, were capable of learning from the context in which words were used and thus provided more accurate sentiment classifications (Schumaker & Chen, 2009). Nevertheless, traditional ML techniques, such as Naive Bayes, support vector machines, and decision trees, often failed to capture long-term dependencies in the text and were susceptible to the high dimensionality of the text data (Chen, Fengler, & Härdle, 2018).

However, deep learning (DL) techniques and their application to NLP has significantly evolved sentiment analysis. This is particularly evident with the emergence of advanced language models (LMs) such as ELMo (Peters et al., 2018), ULMFit (Howard & Ruder, 2018), and transformer-based architectures like BERT, which exhibit superior performance in sentiment classification tasks. With their ability to comprehend context, these models can capture long-term dependencies in text and effectively reduce data dimensionality via high-quality embeddings (Refaeli & Hajek, 2021). Domain-specific adaptations of BERT, like FinBERT, further extend these capabilities, specifically tuned for financial texts, hence bolstering the accuracy and reliability of financial sentiment analysis. In particular, Leippold (2023) emphasized the robustness of FinBERT over keyword-based methods in financial sentiment analysis in manipulated text by adversarial attacks. Farimani, Jahan, Fard, and Tabbakh (2022) showcased the application of FinBERT models for real-time market prediction, emphasizing the importance of sentiment-aware time series derived from financial news related to forex and cryptocurrency markets. Their approach, which combined sentiment scores with technical indicators, demonstrated significant improvements over existing baselines. These developments highlight the transformational potential of LMs in the field, enabling a better understanding and prediction of sentiments.

Despite recent advancements, there are inherent challenges in financial sentiment analysis. These include not only the need to comprehend domain-specific jargon and to disentangle subtle sentiments associated with multiple financial instruments within the same context, but also the ability to adapt the sentiment output based on specific use-cases, prior conditions or topics. For instance, a sentence might be perceived as negative for the general market economy but simultaneously positive for a specific stock. Current models, such as FinBERT, primarily consider the perspective of the experts that annotated the training data (Malo, Sinha, Korhonen, Wallenius, & Takala, 2014), limiting their flexibility to adjust their output based on these finer nuances and making such framing of sentiment context-dependent an open challenge. This highlights the necessity for future research directions that enhance the adaptability and performance of sentiment analysis in finance. In this regard, the advent of LLMs like GPT offers promising potential. The inherent capabilities of these models, including contextual understanding and adaptability, could provide valuable solutions to these challenges, opening new avenues for exploration and study in this domain (Radford et al., 2019).

2.2. ChatGPT and related AI tools

With the success of GPT, an array of LLMs have been developed that showcase remarkable performance in various NLP tasks. Each of these models brings unique advantages and capabilities to the field of NLP, including financial applications.

Among these, BloombergGPT stands out as a prominent example that has exhibited excellent performance in several financial NLP tasks (Wu et al., 2023). Designed by Bloomberg's AI team, this model has been trained on a massive corpus of financial texts. However, despite its notable success in various financial tasks, BloombergGPT is yet to provide an open API, with its application being primarily internal to Bloomberg as of May 2023.

Another noteworthy mention in the realm of LLMs is Google's Bard the main rival of ChatGPT. This AI conversational service is powered by Google's LAMDA (Language Model for Dialogue Applications), combining architectural features of BERT and GPT (Thoppilan et al., 2022). Despite its significant potential for engaging and contextually aware dialogues, it shares a similar limitation with BloombergGPT; as of the time of writing, there is a widely open API available for Bard.

An open-source alternative to GPT-3, BLOOM (Scao et al., 2022), has also gained recognition among prominent LLMs. BLOOM, though open-source, poses its own challenges. To effectively utilize BLOOM, a considerable amount of specialized knowledge is required, alongside sufficient computational resources. Furthermore, there is not a tuned version available for conversation-specific tasks, a feature that makes models like ChatGPT particularly useful.

Post the advent of ChatGPT, numerous LLMs with a focus on more specific tasks have emerged, such as those designed for code completion (Dakhel et al., 2023), content creation, and marketing applications. These models, although more narrow in their application, bring additional dimensions of utility and specialization to the table, further expanding the possibilities and impacts of LLMs. However the remarkable advancement and proliferation of large language models, ChatGPT maintains a leading position within the field (JasperAI, 2023). Its accessibility through an open API, vast scale of training data, and applicability across an array of tasks highlights its considerable potential and practical utility.

Notwithstanding the advancements in the application of ChatGPT across various domains such as healthcare and education (Sallam, 2023), its explicit application to financial sentiment analysis remains an underexplored area. Interestingly, there seems to be a paucity of studies directly evaluating the performance of ChatGPT through its API, which is a prerequisite for broader use and integration into third-party applications. No known studies have, as yet, evaluated the performance of ChatGPT specifically in the domain of financial sentiment analysis, and particularly within the context of the forex market. This presents a remarkable gap in the current body of research that this present study seeks to address. By casting our investigative lens on the application and performance of ChatGPT within the sphere of sentiment analysis, and by directly interacting with its API, we aim to contribute novel insights and findings to this emergent field. This underlines the real-world applicability and practicality of our approach, as it aligns with the deployment scenario of incorporating ChatGPT into applications and services.

3. Methodology

In this section, we describe our methodology, focusing on evaluating ChatGPT's performance in financial sentiment analysis. We discuss data collection, preprocessing, and the deployment of ChatGPT using its API. Details about the experimental setup, including prompt design and evaluation criteria, are also presented, with an emphasis on the practical implications of our findings in the financial sector.

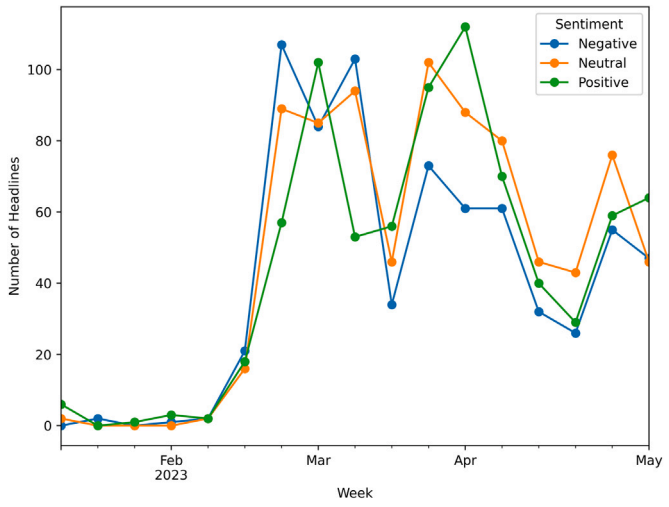

Fig. 1. Time series plot of sentiment labels per week.

3.1. Dataset creation and annotation

For a comprehensive exploration in this study, we collected news headlines relevant to key forex pairs: AUDUSD, EURCHF, EURUSD, GBPUSD, and USDJPY. The data was extracted from reputable platforms Forex Live[1] and FXstreet[2] over a period of 86 days, from January to May 2023 (Fig. 1). It is worth noting that this dataset falls outside the training data of ChatGPT, which concluded in September 2021, ensuring an unbiased evaluation. According to Similarweb (Yeshayahou, 2021), these platforms are recognized as leading sources for reliable news and current forex analyses, offering critical insights into significant economic and political events.

Our dataset comprises 2291 unique news headlines. Each headline includes an associated forex pair, timestamp, source, author, URL, and the corresponding article text. Data collection was accomplished using web scraping techniques executed via a custom service on a virtual machine. This service periodically retrieves the latest news for a specified forex pair (ticker) from each platform, parsing all available information. The collected data is then processed to extract necessary details such as the article's timestamp, author, and URL. The URL is further used to retrieve the full text of each article. Finally, the data are stored in a database (Fig. 2). This data acquisition process repeats approximately every 15 min.

To ensure the reliability of the dataset and the validity of the experimental evaluations, we manually annotated each headline for sentiment. Instead of solely focusing on the textual content, we ascertained sentiment based on the potential short-term impact of the headline on its corresponding forex pair. This method recognizes the currency market's acute sensitivity to economic news, which significantly influences many trading strategies (Evans & Lyons, 2005). As such, this dataset could serve as an invaluable resource for fine-tuning sentiment analysis models in the financial realm.

We used three categories for annotation: 'positive', 'negative', and 'neutral', which correspond to bullish, bearish, and hold sentiments respectively, for the forex pair linked to each headline. Table 1 provides examples of annotated headlines along with brief explanations of the assigned sentiment. Figs. 3 and 4 represent the overall sentiment distribution within the dataset, as well as the distribution for each individual forex pair.

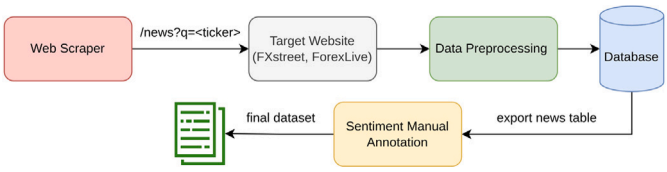

Fig. 2. Dataset creation process.

Furthermore, Figs. 5 and 6 display the most common tokens in the headlines categorized as positive and negative sentiment, after stopwords, forex pairs, and central banks' names have been removed. It is important to note that the classification of tokens as positive or negative is based on the overall sentiment of the headlines in which they appear. Specifically, tokens found in headlines labeled as positive are displayed in Fig. 5, while those in headlines labeled as negative are shown in Fig. 6. The tokens themselves were not individually classified as positive or negative; instead, their classification was derived from the sentiment of the headlines they were part of.

Table 2 provides further quantitative detail on our dataset, including the distribution of tokens, the number of articles, and the mean and standard deviation of daily articles for each forex pair. The total number of articles collected for each pair ranges from 55 for EURCHF to 758 for EURUSD. Specifically, AUDUSD has 461 articles with an average of 5.36 articles daily, GBPUSD comprises 518 articles averaging 6.02 daily, and USDJPY includes 499 articles with a daily average of 5.8. The daily average articles vary for each pair, with EURUSD having the highest daily average of 8.81 articles and EURCHF recording the lowest with 0.64 articles per day.

Furthermore, we identified the number of missing dates within the total period of 113 dates for each forex pair. AUDUSD has 12 missing dates, EURCHF has 80 missing dates, EURUSD has 16 missing dates, GBPUSD has 18 missing dates, and USDJPY has 24 missing dates. These correspond to the days without any news articles published for each respective ticker during the specified period.

Regarding the text itself, the average number of tokens in each article text is approximately 263, with a standard deviation of 159. For headlines, the average number of tokens is around 12, with a standard deviation of 3. These statistics illustrate the variability in text and headline lengths within the dataset.

Importantly, the number of tokens plays a significant role in the response latency, cost, and performance of language models like ChatGPT. Longer texts require more computational resources and time for processing, which can lead to increased latency and cost. Conversely, shorter texts can be processed faster and at a lower cost, but they might not provide as much contextual information for the model to generate accurate and relevant responses.

In an effort to contribute to the research community and foster transparency in our methodology, we have made our curated dataset publicly accessible. This collection, featuring news headlines linked to major forex pairs alongside their respective sentiments, is among the first of its kind. We anticipate that it will serve as an invaluable resource for researchers and practitioners keen on exploring or applying machine learning techniques, specifically within the scope of financial sentiment analysis.

We hope that this contribution will spur further research in this field, enriching our shared comprehension of the dynamic interplay between financial news sentiment, AI models, and market behavior. Detailed information about how to access and use the dataset is available at Zenodo[3] repository (Fatouros & Kouroumali, 2023).

[1] https://www.forexlive.com/.
[2] https://www.fxstreet.com/.
[3] https://zenodo.org/record/7976208.

Exhibit 1013 Page 4

Table 1
Examples of annotated headlines.

Forex pair	Headline	Sentiment	Explanation
GBPUSD	Diminishing bets for a move to 12 400	Neutral	Lack of strong sentiment in either direction
GBPUSD	No reasons to dislike Cable in the very near term as long as the Dollar momentum remains soft	Positive	Positive sentiment towards GBPUSD (Cable) in the near term
GBPUSD	When are the UK jobs and how could they affect GBPUSD	Neutral	Poses a question and does not express a clear sentiment
USDJPY	BoJ's Ueda: Appropriate to continue monetary easing to achieve 2% inflation target with wage growth	Positive	Monetary easing from Bank of Japan (BoJ) could lead to a weaker JPY in the short term due to increased money supply
USDJPY	Dollar rebounds despite US data. Yen gains amid lower yields	Neutral	Since both the USD and JPY are gaining, the effects on the USDJPY forex pair might offset each other
USDJPY	USDJPY to reach 124 by Q4 as the likelihood of a BoJ policy shift should accelerate Yen gains	Negative	USDJPY is expected to reach a lower value, with the USD losing value against the JPY.
AUDUSD	RBA Governor Lowe's Testimony High inflation is damaging and corrosive	Positive	Reserve Bank of Australia (RBA) expresses concerns about inflation. Typically, central banks combat high inflation with higher interest rates, which could strengthen AUD.

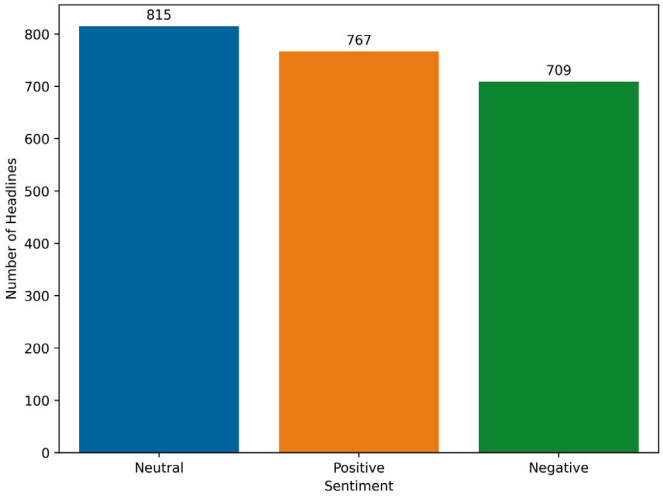

Fig. 3. Sentiment distribution.

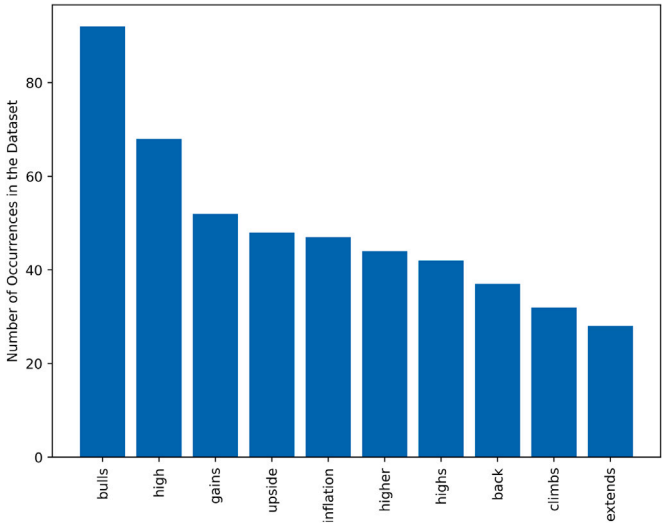

Fig. 5. Most common tokens on positive headlines.

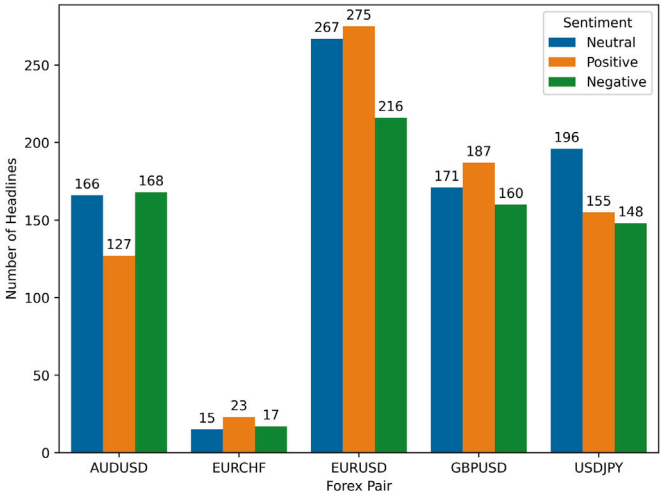

Fig. 4. Sentiment distribution per forex pair.

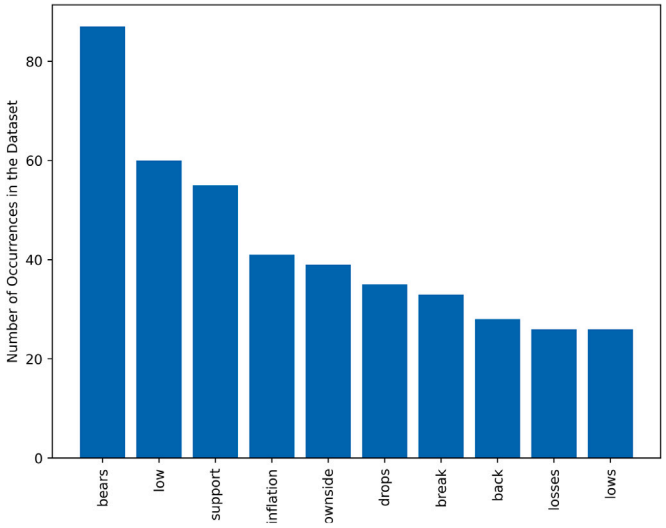

Fig. 6. Most common tokens on negative headlines.

3.2. Establishing baseline: Sentiment classification using FinBERT

To validate the effectiveness of our proposed approach, we contrast ChatGPT's performance with a conventional benchmark in financial sentiment analysis — FinBERT. Selected for its exceptional proficiency in classifying sentiments expressed in financial texts, FinBERT provides a robust baseline for comparing and evaluating our results. By doing so, we aim to demonstrate the substantial potential of ChatGPT in this specialized task.

In terms of implementation, we utilized the Transformers library from Hugging Face to operate FinBERT (Wolf et al., 2020). This Python-based API provides an intuitive interface for employing pre-trained transformer models. By leveraging the convenience and ease of use of the Transformers library, we efficiently integrated FinBERT into our experiments. This high-level process involved loading the FinBERT model and tokenizer from Hugging Face's model repository, tokenizing the headlines, and feeding them into the model. The FinBERT model subsequently outputs a set of probabilities for each sentiment class (positive, negative, and neutral), representing the model's confidence in associating the input headline with each sentiment category. These probabilities are used to determine the predicted class and a sentiment score (notated as FinBERT-N) for each headline. The sentiment score is computed by subtracting the negative class probability from the positive class probability.

3.3. Exploring sentiment classification with ChatGPT

The primary goal of this work is to explore the potential of ChatGPT in the field of financial sentiment analysis. Given its extensive training on diverse texts, ChatGPT is expected to demonstrate an advanced understanding of context, a key factor in interpreting financial news sentiment. The rationale for utilizing ChatGPT lies in its potential ability to comprehend not only the literal meanings of words but also the underlying implications, idioms, or sentiments that can be industry-specific or topic-centric. For example, in the forex market, a piece of news might have a positive sentiment for one currency (e.g., JPY) but a negative sentiment for the corresponding forex pair (e.g., USDJPY). This nuanced contextual understanding, crucial to most sentiment or emotion analysis tasks, is where ChatGPT could potentially outshine established LMs such as BERT.

To evaluate the capabilities of ChatGPT in financial sentiment analysis, we devised and conducted a series of experiments through a zero-shot prompting approach. With zero-shot prompting, we harness ChatGPT's pre-existing training and understanding of language patterns and context to generate desired responses without any task-specific fine-tuning or further training. This approach allows us to directly evaluate ChatGPT's inherent capabilities in sentiment analysis without any additional modifications, demonstrating the potential for direct integration and utilization of ChatGPT in various applications.

Table 3 details the prompts used for sentiment classification. These prompts range from emulating the perspectives of a financial analyst evaluating news (GPT-P1 and P2) to a sentiment analysis AI model (such as FinBERT) assessing a headline (GPT-P3), a forex trader reacting to market updates (GPT-P4), as well as a prompt evaluating all the available daily news per forex pair (GPT-P5) and all available daily news (GPT-P6). Furthermore, we implemented variations of these prompts requiring numerical sentiment output in the form of a sentiment score (GPT P1N-P6N) ranging from $[-1, 1]$ instead of a sentiment class label.

An additional variant, applied to FinBERT and GPT-P4 (FinBERT-A, FinBERT-AN, GPT-P4 A, and GPT-P4AN), was set up to analyze model performance in sentiment analysis when both the article text and news headline are supplied. This aims to determine whether the inclusion of the article's text – as additional context – enhances the model's accuracy.

Table 2
Dataset statistics.

Forex pair	No of articles	Daily articles	Headline Tokens	Article Tokens
AUDUSD	461	5.36 (3.97)	11.51 (2.85)	288.42 (149.4)
EURCHF	55	0.64 (0.92)	12.58 (3.06)	162.62 (98.9)
EURUSD	758	8.81 (7.08)	11.56 (2.9)	255.87 (180.53)
GBPUSD	518	6.02 (4.67)	11.95 (2.55)	263.1 (148.29)
USDJPY	499	5.8 (4.29)	11.68 (2.88)	259.38 (142.82)
Total	2291	26.64 (21.71)	11.69 (2.82)	262.58 (158.99)

Note: Values in parenthesis represent standard deviations.

Specifically, prompts GPT-P6 and GPT-P6N were crafted to prompt ChatGPT to process a collection of headlines and yield a cumulative sentiment in a structured JSON format. Evaluating this feature is key to understanding ChatGPT's adaptability and its potential for seamless integration into existing platforms. This capability is particularly essential for practical scenarios where the generated outputs are directly fed into other systems or services.

Prompts GPT-P6 and P6N were designed to ask ChatGPT to process a list of headlines and return a collective sentiment in a structured format, specifically JSON. This feature evaluates ChatGPT's capability to integrate seamlessly into existing services by producing easily consumable output. Such a capability is vital in real-world applications where the model's results need to be immediately utilized by other services or systems.

To conduct these sentiment analysis experiments, we leveraged the GPT-3.5-turbo model available from OpenAI's Python library (Brown et al., 2020). Each prompt was fed to the model as a single user instruction. Depending on the prompt, we fine-tuned several parameters to optimize the model's output. Specifically, the *max_tokens* parameter in GPT P1-P5 was set to 1 to constrain the model's response to a single token. For prompts GPT P1N-P4N, P5N, P6-P6N, the *max_tokens* parameter was set to 10, 20, and 200 tokens respectively, according to their expected response length. Consequently, dedicated processing was performed on the response text to extract the necessary information per prompt. For instance, we crafted a straightforward function with a regular expression to extract the JSON (JavaScript Object Notation) from P6's response. The *temperature* parameter was set to 0.2 to guide the model towards generating more deterministic outputs. It is worth noting that we ended up with these prompts after extensive prompt engineering to ensure that GPT includes the desired information (i.e., sentiment class or score) in its responses.

For each prompt, we monitored the time taken to receive a response from the API to track the task latency and evaluate GPT's efficiency. We also logged the number of tokens used in the completion and the number of tokens in the prompt for cost management and resource usage tracking.

Each experiment was run iteratively over the forex dataset presented in Section 3.1. To ensure the smooth execution of the iterative process, we incorporated error handling mechanisms to address potential API exceptions or response errors. It is essential to note that running these sentiment analysis experiments over large datasets consumes API tokens and incurs costs in line with OpenAI's pricing structure. However, the insights obtained from the sentiment analysis greatly outweigh the computational expenses, offering invaluable intelligence for enhanced decision making in finance.

3.4. Evaluation metrics

The evaluation approach for sentiment classification in this study was twofold, encompassing both a traditional evaluation grounded in

Table 3
Experimental ChatGPT prompts for sentiment classification.

Prompt Abbr.	Prompt
GPT-P1	Act as a financial expert holding {ticker}. How do you feel about the headline {headline}? Answer in one token: positive, negative, or neutral.
GPT-P2	Act as a financial expert. Classify the sentiment for {ticker} based only on the headline {headline}. Answer in one token: positive, negative, or neutral.
GPT-P3	Act as a sentiment analysis model trained on financial news headlines. Classify the sentiment of the headline {headline}. Answer in one token: positive, negative, or neutral.
GPT-P4	Act as an expert at forex trading holding {ticker}. Based only on the headline {headline}, will you buy, sell or hold {ticker} in the short term? Answer in one token: positive for buy, negative for sell, or neutral for hold position.
GPT-P5	Act as an expert at forex trading holding {ticker}. Based only on the following list of headlines {ticker_daily_headlines}, will you buy, sell or hold {ticker} in the short term? Answer in one token: positive for buy, negative for sell, or neutral for hold position
GPT-P6	Act as a sentiment analysis service of a financial platform. Based only on the following list of headlines {all_daily_headlines}, provide a summary of the daily sentiment for the forex pairs: {tickers}. Provide only the sentiment per forex pair in JSON format eg {'USDJPY': 'positive', 'EURUSD': 'neutral'}. The sentiment can be positive for buy, negative for sell, or neutral for hold position.

the comparison with the true sentiment class and a market-related model evaluation. The former method focused on the models that predicted a sentiment class for each headline (i.e., FinBERT and GPT's P1-P4), by comparing the predicted sentiment class with the actual sentiment class as annotated in our dataset. On the other hand, the market-related model evaluation was applicable to all considered models, achieved by assigning the corresponding integer code to the predicted sentiment classes. Specifically for FinBERT, the sentiment score was derived from the predicted probabilities per class, as detailed in Section 3.2.

3.4.1. Evaluation based on true class

To gauge the effectiveness of our sentiment classification models, we devised an evaluation framework rooted in multiple performance metrics, such as accuracy, precision, recall, and F1-score. These metrics illuminate the extent to which the models accurately predict sentiment classes defined in our annotated dataset.

A key metric in our evaluation toolbox is the Sentiment Mean Absolute Error (S-MAE) expressed by Eq. (1), where N represents the total number of instances or data points, y_i denotes the true sentiment class for the ith instance, and $\hat{y}_i$ represents the predicted sentiment class for the ith instance. This metric quantifies the mean absolute difference between the integer-represented true sentiment classes and the predicted sentiment classes, with -1, 0, and 1 representing negative, neutral, and positive sentiments, respectively.

$$S-MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{1}$$

The rationale behind utilizing S-MAE is its capacity to distinctly penalize models that significantly misjudge sentiment classification. This aspect is particularly important given the intricate nature of sentiment analysis in financial news, where sentiment interpretation can be complex and nuanced. For instance, a headline with a positive sentiment towards a particular currency (e.g., USD) could suggest a negative sentiment for its paired currency (e.g., EURUSD). In another case, a negative sentiment towards the economy due to an announcement from a central bank might convey a positive market sentiment for the corresponding currency. These subtleties underscore the necessity of penalizing not only incorrect polarity predictions but also off-target intensity estimations.

Moreover, the evaluation includes a comparative analysis among different models. Specifically, we compare the performance of the FinBERT model with that of ChatGPT under varying prompts (P1-P4). This approach intends to highlight the impact of prompt selection on the performance of ChatGPT, offering a nuanced understanding of how these different models perform in sentiment classification tasks.

3.4.2. Model evaluation in relation to the market

Aside from conventional classification metrics, we also evaluated the models' output in connection with actual market returns. This evaluation strategy is driven by the primary intended application of these models — to guide trading decisions in financial markets.

For this component of the evaluation, we aggregate the sentiment scores for each forex pair on a daily basis. In particular, the sentiment scores of all headlines concerning a specific forex pair within a single day are summed to derive a daily sentiment score (see Eq. (2)). This score is then compared with the actual daily return of the corresponding forex pair using Pearson correlation. This correlation is computed to assess whether the sentiment scores, as predicted by the models, have any relation to real market movements.

$$S_d = \sum_{i=1}^{N_d} s_i \tag{2}$$

where:

- S_d is the overall sentiment score for day d.
- N_d is the number of headlines related to the forex pair on day d.
- s_i is the sentiment score of the ith headline on day d.

In addition to correlation with market returns, the models are also evaluated on their ability to accurately predict the direction of market movement, a critical element in financial decision-making (Blaskowitz & Herwartz, 2011). This metric, termed Directional Accuracy (DA), is computed as the percentage of days where the predicted sentiment direction (positive/negative) aligns with the actual market return direction. For example, if the model correctly predicted the direction of market movement on 65 days out of 100, then the DA of the model would be 65%.

For the purpose of this market-related evaluation, the sentiment score for prompts GPT-P1 to P6 is regarded as an integer value of the predicted class. This interpretation allows the sentiment score to be quantified and aggregated in a way that aligns with its intended meaning — a measure of the overall daily sentiment towards a specific forex pair.

To provide a comprehensive benchmark, we calculated both a naive baseline for DA and the DA based on the true sentiment. The naive approach simply predicts the most frequent sentiment direction for each forex pair in the annotated dataset. Given the imbalances often present in real-world data, such a baseline offers a foundational comparison point. Meanwhile, the DA of the true sentiment serves as a direct measure of the inherent alignment between sentiment and market movement in our dataset.

This evaluation includes all models (FinBERT, GPT P1-P6, and P1N-P6N) and, by combining DA with Pearson correlation, offers a more holistic view of the models' performance in a financial context.

This evaluation includes all models (FinBERT, GPT P1-P6, and P1N-P6N) and, by combining Directional Accuracy with Pearson correlation, offers a more holistic view of the models' performance in a financial context. The results underscore the importance of sophisticated sentiment analysis techniques, especially when compared to the naive baseline, in capturing the nuances of financial market movements.

Exhibit 1013 Page 7

Table 4
Performance results in sentiment classification.

Model	Accuracy	Precision	Recall	F1	S-MAE
FinBERT	0.561	0.560	0.562	0.556	0.540
GPT-P1	0.730	0.760	0.730	0.725	0.300
GPT-P2	**0.790**	0.797	**0.790**	**0.790**	0.227
GPT-P3	0.735	0.780	0.735	0.737	0.282
GPT-P4	0.784	**0.804**	0.784	0.789	**0.221**

4. Results

This section details the findings obtained from our extensive analysis and experiments, aimed at scrutinizing the effectiveness of ChatGPT in financial sentiment analysis. We present a comprehensive discussion of these results, comparing the performance of the various ChatGPT prompts and the established baseline, FinBERT, across various metrics. The code for reproducing the presented results is available at GitHub.[4]

4.1. Sentiment classification performance

Table 4 showcases the sentiment classification performance of various models including FinBERT and distinct prompts from the ChatGPT (GPT-P1 to P4) used in our dataset. The evaluation metrics include Accuracy, Precision, Recall, F1-Score, and S-MAE.

As presented in the table, GPT models consistently outperform FinBERT across all metrics. FinBERT, which is a general-purpose language model trained on financial text, shows an accuracy of 0.561, a precision of 0.560, and a recall of 0.562. Its F1-Score stands at 0.556 and it yields an S-MAE of 0.540. Although these figures are decent, they are significantly overshadowed by the performance of the GPT models. Interestingly, GPT-P3, which mirrors the FinBERT model in acting as a sentiment analysis model trained on financial news headlines performs relatively well. However, it does not match the performance of other GPT models that leverage the forex pair information in their prompts.

GPT-P1, which is designed to function as a financial expert performing emotion analysis on headlines while taking into account the related forex pair, also registers a commendable performance in the sentiment classification task. However, it underperforms slightly compared to GPT-P2, with its metrics trailing by approximately 8%, indicating that while the emotion-based approach of P1 is effective, the sentiment-oriented approach of P2 appears to offer more precision in this context.

GPT-P2, and P4, which act as financial experts considering the related forex pair in their sentiment analysis, register substantial improvements in the sentiment classification task. GPT-P2 leads the pack with the highest accuracy, recall, and F1-Score around 0.790. GPT-P4, designed to impersonate a forex trader making trading decisions based on the headline sentiment, posts similar performance metrics to GPT-P2, despite the different framing of the prompt. This aligns with how the dataset was annotated, indicating that role-playing prompts, such as those used in GPT-P4 and GPT-P2, are especially effective.

Table 5 offers a more granular understanding of the models' performance, further evaluating their precision, recall, and F1-score for each individual sentiment class: Positive, Negative, and Neutral. For the Positive sentiment class, GPT-P1 exhibited the highest precision, while GPT-P4 led in terms of recall and F1-score. In the Negative sentiment class, GPT-P4 showcased the highest precision, while GPT-P2 dominated in recall and F1-score. For the Neutral sentiment class, GPT-P2 outperformed in precision, while GPT-P3 achieved the highest recall. However, both GPT-P2 and GPT-P4 tied for the top spot in F1-score for Neutral sentiment. These detailed metrics offer insights into the strengths and weaknesses for each model across different sentiment classes. Such a breakdown is crucial for applications where the accurate classification of a specific sentiment class is more critical than overall performance.

Table 5
Performance metrics for individual sentiment classes across models.

	FinBert	P1	P2	P3	P4
Positive					
Precision	0.59	**0.91**	0.87	0.89	0.88
Recall	0.61	0.54	0.71	0.59	**0.79**
F1-score	0.60	0.68	0.79	0.71	**0.82**
Negative					
Precision	0.56	0.74	0.81	0.87	**0.88**
Recall	0.68	**0.90**	0.88	0.72	0.76
F1-score	0.61	0.81	**0.85**	0.79	0.81
Neutral					
Precision	0.53	0.64	**0.71**	0.60	0.66
Recall	0.41	0.77	0.78	**0.88**	0.83
F1-score	0.46	0.69	**0.74**	0.72	**0.74**

Table 6
Best performing model in sentiment classification per forex pair.

FX pair	Accuracy	Precision	Recall	F1	S-MAE
AUDUSD	P2	P4	P2	P2	P2
EURCHF	P2	P4	P2	P2	P4
EURUSD	P4	P4	P4	P4	P4
GBPUSD	P4	P4	P4	P4	P4
USDJPY	P2	P2	P2	P2	P2

Furthermore, as depicted in Table 6, the performance of the GPT models varies depending on the specific forex pair. For the AUDUSD pair, GPT-P2 demonstrates superior performance in terms of accuracy, recall, F1-score, and S-MAE, while GPT-P4 outperforms the other models in precision. In the case of the USDJPY pair, GPT-P2 dominates all performance metrics. The evaluation of the EURCHF pair reveals a split performance dominance, with GPT-P2 leading in accuracy, recall, and F1-score, while GPT-P4 stands out in precision and S-MAE. Notably, GPT-P4 dominates in all performance metrics for both the EURUSD and GBPUSD pairs, underlining the effectiveness of GPT-P4 for sentiment analysis specifically associated with these forex pairs.

However, it is worth noting that 75.6% of the headlines in the dataset mention the related forex pair at their beginning. This observation suggests that the performance difference between prompts such as GPT-P1, P2, and P4, which take the forex pair into account, could be even more pronounced when dealing with more complex headlines. To explore this, we further examined the performance of the models specifically on a subset of the data with headlines that do not explicitly mention the associated forex pair. This assessment provides insight into how well each model handles sentiment analysis in a more challenging context, where the pertinent topic (i.e., forex pair) is not directly stated in the headline text. The results from this analysis, presented in Table 7, showed that GPT-P4 continued to outperform, achieving the highest position across all metrics. This finding underscores GPT-P4's ability to interpret the implicit context in the headlines.

On the other hand, FinBERT's performance slightly declined when the forex pair was not mentioned directly, with accuracy, precision, recall, F1-score, and S-MAE at 0.543, 0.543, 0.543, 0.538, and 0.539 respectively. GPT-P1, P2, and P3 also experienced a dip in performance compared to the results from the complete dataset. However, even with the drop in performance, GPT-P1, P2, and P3 outperformed FinBERT, underlining the superiority of the GPT models in handling complex scenarios. In the case of GPT-P2, despite a decrease in performance, it still secured high scores with accuracy, precision, recall, F1-score, and S-MAE at 0.711, 0.728, 0.711, 0.714, and 0.326 respectively, emphasizing its overall robustness in sentiment classification.

[4] https://github.com/giorgosfatouros/Financial-Sentiment-Analysis-with-ChatGPT.

Exhibit 1013 Page 8

Table 7
Performance results in sentiment classification (filtered data).

Model	Accuracy	Precision	Recall	F1	S-MAE
FinBERT	0.543	0.543	0.543	0.538	0.539
GPT-P1	0.663	0.715	0.663	0.672	0.410
GPT-P2	0.711	0.728	0.711	0.714	0.326
GPT-P3	0.665	0.657	0.665	0.657	0.384
GPT-P4	**0.765**	**0.772**	**0.765**	**0.763**	**0.249**

Table 8
Correlation of predicted sentiment and forex pair returns.

Model	AUD/ USD	EUR/ CHF	EUR/ USD	GBP/ USD	USD/ JPY
FinBERT	0.004	−0.214	0.192	0.206	0.209
FinBERT-N	−0.206	−0.028	0.155	0.172	0.113
FinBERT-A	0.282	−0.134	0.169	0.051	0.448
FinBERT-AN	0.313	−0.220	0.122	0.085	0.505
GPT-P1	0.401	−0.025	0.461	0.398	0.659
GPT-P1N	0.353	**0.235**	0.547	0.461	0.627
GPT-P2	0.413	−0.060	0.446	0.441	0.659
GPT-P2N	0.317	−0.061	0.476	0.438	0.606
GPT-P3	0.405	−0.118	0.440	0.460	0.619
GPT-P3N	0.390	0.035	0.464	0.451	0.650
GPT-P4	0.508	0.061	0.519	0.526	0.641
GPT-P4N	0.466	0.058	0.388	0.414	0.542
GPT-P4A	0.499	0.004	0.473	0.335	0.646
GPT-P4AN	0.532	0.0	0.466	0.342	0.581
GPT-P5	0.132	0.131	0.428	0.335	**0.669**
GPT-P5N	0.335	0.132	0.416	0.404	0.448
GPT-P6	0.580	0.072	0.461	0.527	0.628
GPT-P6N	**0.589**	−0.098	**0.585**	**0.557**	0.620
True Sent	0.517	0.173	0.425	0.496	0.650

This analysis highlights the strength of GPT models, particularly GPT-P4, in sentiment analysis tasks. These models' ability to maintain a high level of performance even in more challenging scenarios makes them promising tools for real-world applications. Crucially, these results also emphasize the importance of the prompt in instructing ChatGPT. By thoughtfully crafting prompts based on the specifics of the task at hand, we can guide the model to leverage its knowledge more effectively, thus enhancing its performance in sentiment analysis.

4.2. Sentiment score relation to the financial market

The second evaluation approach examines how well the sentiment scores predicted by each model align with market price movements. We compare the sentiment scores of each model against the actual daily returns of the respective forex pairs to understand the potential utility of the models in a trading context. The calculation of the overall daily sentiment score is detailed in Section 3.4.2. This comparison was done in terms of Pearson correlation and DA as presented in Section 3.

4.2.1. Correlation with market returns

The correlation coefficient between sentiment scores and market returns is a key factor in evaluating the effectiveness of the presented models. This correlation is measured using the Pearson correlation coefficient, which ranges from −1 to +1. The higher the correlation value, the more directly related the sentiment scores and market returns are. In this context, a higher positive correlation implies that as sentiment improves, market returns also increase. Consequently, the model with the highest correlation between its sentiment scores and market returns can be considered the most accurate in reflecting the market's response to sentiment changes.

The correlation matrix represented in Fig. 7 unveils the interplay among various models, the annotated sentiment, and the daily market returns calculated applying pairwise correlation to our dataset aggregated on a daily basis as discussed in Section 3.4.2. It is indeed reasonable that GPT-P4 exhibits the highest correlation with the true

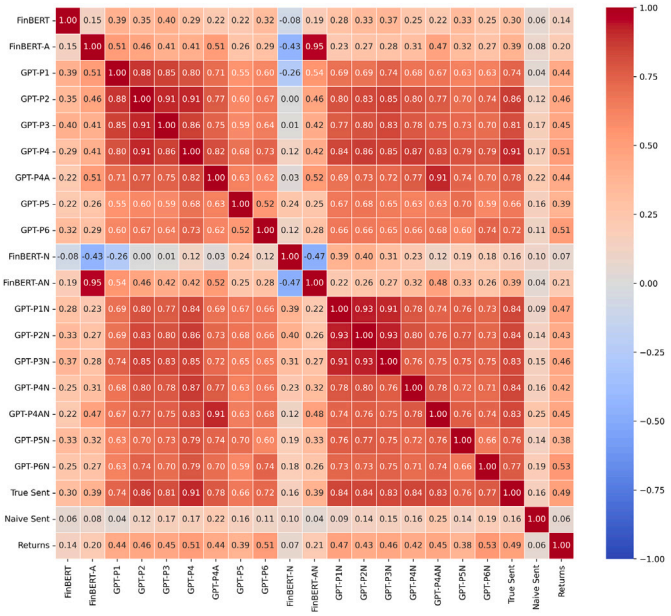

Fig. 7. Correlation matrix of predicted sentiment and market returns.

sentiment, given that P4 mirrors the approach used in annotating the sentiment within the dataset. Similarly, FinBERT correlates the most with GPT-P3 which mimics a sentiment analysis model trained on financial data. However, a more intriguing observation is that GPT-P6N demonstrates a higher correlation with the market returns than the true sentiment itself. This underscores the power of GPT's ability to process all daily news simultaneously, which enhances its capacity to gauge market sentiment with a higher degree of accuracy. Most models generating sentiment scores ranging from −1 to 1 instead of performing sentiment classification exhibit a more fitting alignment with the market movements. This is plausible as the latter models are better positioned to accurately capture news sentiment that is slightly positive or negative.

Interestingly, while the GPT models considering entire article texts in their context (specifically, GPT-P4 A and GPT-P4AN) did not outperformed their counterparts that focus solely on headlines, the opposite holds true for FinBERT. It exhibited enhanced performance when presented with a more expansive context.

It is also worth noting that GPT-P5 showed the lowest correlation with both daily returns and true sentiment, indicating that this prompt may not be as effective in sentiment prediction for market price prediction as the others.

Furthermore, the high correlation between GPT-P1 and P1N, through to GPT-P6 and P6N, suggests that the different versions of the models (categorical and numerical (N) versions) largely agree on the sentiment of the text. This indicates the consistency of the GPT's outputs.

The correlation between true sentiment and daily returns stands at around 0.49, implying a moderate linear relationship. This suggests that sentiment can explain some but not all of the variations in market prices. Other factors and market dynamics likely come into play as well.

Delving into the specifics, Table 8 elaborates on the correlation coefficients between predicted sentiments by various models and returns for individual forex pairs. Though different models manifest diverse correlation strengths with separate currency pairs, the pair-specific correlations for each model largely echo the trends spotlighted in Fig. 7. Nonetheless, a consistent observation across all models is the low or negative correlation with EURCHF pair returns, which can likely be attributed to the limited volume and scarcity of news concerning the EURCHF.

Exhibit 1013 Page 9

Table 9
Directional accuracy of non-numerical and naive models.

Model	DA
FinBERT	0.595
FinBERT-A	0.498
GPT-P1	0.583
GPT-P2	0.607
GPT-P3	0.607
GPT-P4	0.640
GPT-P4A	0.587
GPT-P5	**0.652**
GPT-P6	0.640
Naive Sent.	0.521

Table 10
Directional accuracy of numerical and true sent. models.

Model	DA
FinBERT-N	0.599
FinBERT-AN	0.526
GPT-P1N	**0.672**
GPT-P2N	0.644
GPT-P3N	0.660
GPT-P4N	0.611
GPT-P4AN	0.648
GPT-P5N	0.623
GPT-P6N	0.652
True Sent.	0.696

4.2.2. Directional accuracy of sentiment scores

Tables 9 and 10 present the DA of different sentiment scoring models, as well as the naive baseline and the DA based on the true sentiment. DA reflects the percentage of instances where the predicted and actual direction of change align, with regards to sentiment scores and market movements.

For context, the naive approach, which predicts the most frequent sentiment direction for each forex pair, achieves a DA of 52.1%. This serves as a baseline for our evaluations. In comparison, the DA based on the actual sentiment stands at a notably higher 69.6%.

Examining the tables, GPT-P1N is found to have the highest DA at 67.2%, indicating the superior ability of this model to predict the correct direction of market movement based on sentiment scores. Close contenders are GPT-P3N and GPT-P6N, with DA values around 66%. Among categorical models, GPT-P5 shows the highest DA, closely followed by GPT-P6 and GPT-P4. However, the numerical models, in general, tend to have a higher DA than their categorical counterparts, underlining their enhanced capacity in capturing the directional changes in market sentiment. Remarkably, the GPT models exhibit performance that is nearly on par with the human-annotated sentiment, underscoring their robustness in this task.

Table 11 showcases the DA of numerical sentiment analysis models, broken down per forex pair. From the results, it is evident that models GPT-P1N through GPT-P3N display a remarkable consistency in their performance across all the forex pairs. This uniformity indicates that despite the varied instructions in their prompts, these models capture similar sentiment elements from the headlines.

GPT-P4AN, which incorporates the entire article content, emerges as the top performer for the AUDUSD. However, for the rest pairs it exhibits comparable performance with the other GPT models that utilize only the headlines.

Focusing on the EURUSD and GBPUSD pairs, GPT-P6N distinctly outperforms its counterparts. This could be attributed to the model's ability to process a comprehensive list of daily headlines, allowing it to analyze the collective sentiment more effectively, which in turn might be more reflective of the market movements for these pairs.

Notably, for the EURCHF currency pair, the DA is generally lower. This might be due to considerable less data being available for these pair, reinforcing the critical role of ample data availability in accurate market direction prediction. Although, FinBERT, when considering the entire article content, achieves the highest DA for EURCHF, the low number of data samples does not allows for safe conclusions on the importance of the broader context.

Table 11
Directional accuracy results for each numerical model per ticker.

Model	AUDUSD	EURCHF	EURUSD	GBPUSD	USDJPY
FinBERT	0.537	0.469	0.538	0.547	0.482
FinBERTA	0.481	**0.562**	0.481	0.491	0.625
GPT-P1	0.648	0.500	0.673	0.679	**0.696**
GPT-P2	0.648	0.438	0.577	0.679	0.679
GPT-P3	0.630	0.500	0.635	0.755	0.679
GPT-P4	0.537	0.188	0.500	0.660	0.536
GPT-P4A	**0.667**	0.312	0.615	0.660	0.679
GPT-P5	0.593	0.469	0.615	0.660	0.589
GPT-P6	0.593	0.344	**0.731**	**0.792**	0.607

Note: The results refer to the numerical (N) versions of the models providing a continuous sentiment value in [−1,1].

Table 12
Average time and tokens processed per ChatGPT prompt.

Prompt	Pr. time	Pr. tokens	Hdln. time	Hdln. tokens
P1	0.94	63.0	0.94	63.17
P2	0.81	67.0	0.81	67.15
P3	0.97	66.0	0.97	65.91
P4	0.82	84.0	0.82	84.39
P4A	0.66	437.0	0.82	437.09
P5	0.83	193.0	0.11	24.74
P6	5.99	592.0	0.22	22.22
P1N	2.29	90.0	2.29	90.34
P2N	1.42	85.0	1.42	84.79
P3N	1.23	80.0	1.23	79.98
P4N	1.06	91.0	1.06	90.81
P4AN	0.71	494.0	1.06	494.09
P5N	1.85	201.0	0.24	25.68
P6N	6.10	600.0	0.23	22.54

In essence, these findings underscore the intricate relationship between news sentiment and market movements and emphasize the necessity of comprehensive testing on diverse datasets for model validation. They also highlight the importance of understanding the specific dynamics and trends of each market and tailoring sentiment analysis approaches accordingly.

4.2.3. Performance and cost analysis

In the context of deploying ML models, particularly in a production environment, performance and cost considerations are crucial. These factors can significantly affect the scalability, efficiency, and overall feasibility of a model. Given this, we evaluate the GPT prompts in terms of processing time and token consumption, both of which directly impact operational costs and performance. Table 12 provides a summary of the average processing time and the number of tokens processed by each prompt.

Among the sentiment classification prompts, P1, P2, and P3 exhibit similar average times and tokens. They produce concise responses in less than a second with around 65 tokens. P4, which is the best performing prompt, generates more tokens, which can potentially lead to higher costs. On the other hand, P5 and P6, while generating significantly more tokens, benefit from processing multiple headlines at once, as they have lower average times and tokens per headline, which in addition to higher accuracy, can lead to a cost reduction when processing large quantities of data. The prompts generating numerical outputs follow a similar pattern, but they generally require more time and tokens due to the complexity of generating more accurate responses. Furthermore, the experimental results suggest that the added costs, which can be more than five times the tokens per prompt when analyzing the entire news article, do not translate to a commensurate improvement in performance.

Exhibit 1013 Page 10

It is important to note that the processing times mentioned in the study are dependent on many factors and may not be entirely consistent. The load on the ChatGPT model, which can vary depending on the time of day, other global usage patterns, and end-user location, may significantly impact the processing time. Thus, the average time taken by the models, as presented in the results, while informative for relative comparisons, might not always reflect their real-time performance. Consequently, the time-related results should be interpreted with caution as they might not be entirely reliable.

Taking into account that the cost of utilizing the OpenAI API for the ChatGPT-3.5 model is 0.002 USD per 1 K tokens, the financial implications of integrating ChatGPT into existing services seem relatively low, especially considering the volume of data processed daily. As a reference, Bloomberg News produces approximately 5000 articles on a daily basis (Bloomberg, 2023).

To illustrate, let us consider using the P6N model, which generates the highest number of tokens among the models tested in this study. With this model, the daily cost can be calculated as follows: 600 tokens (average generated by P6N per article) $*$ 5000 articles $*$ 0.002 USD per 1 K tokens. This calculation results in a total daily cost of approximately 6 USD.

Hence, even when dealing with substantial data volumes, such as those reported by Bloomberg, the cost of using advanced language models like ChatGPT for sentiment analysis remains reasonably affordable. This affordability, in combination with their demonstrated performance, makes these models a compelling option for real-world financial applications.

However, for applications that require real-time insights and increased accuracy, a mixed approach combining single and multiple news analysis might be necessary. This would allow the flexibility of generating quick and concise responses for individual headlines with P1-P4 (or P1N-P4N), while also processing a larger batch of headlines with P5, P6 (or P5N, P6N) when time allows.

5. Discussion

5.1. Performance of ChatGPT in financial sentiment analysis

The results of our study reveal that the ChatGPT model consistently outperforms the FinBERT model in financial sentiment analysis, regardless of the specific prompt used. In particular, prompts GPT-P4, P6, P6N showed considerable performance in sentiment analysis, providing potentially valuable insights for predicting market behavior. This high performance, combined with the model's flexibility through prompt engineering, implies significant potential for fine-tuning and optimizing the application. For example, through strategic framing of the task at hand, as demonstrated by prompts GPT4-P4 and P6, ChatGPT's accuracy could be further improved. Despite these promising results, it is important to note that the effectiveness of these prompts varies, which highlights the potential for further optimization and underlines the importance of testing on larger datasets.

Interestingly, GPT-P6N demonstrated a higher correlation with market returns than the actual sentiment, which hints at the model's superior understanding of overall market sentiment when processing all daily news at once. This could potentially be due to ChatGPT's ability to capture more nuanced sentiment variations across a set of news articles compared to individual human annotators who are assessing texts independently.

Moreover, models that generate sentiment scores ranging from -1 to 1 (N versions) showed a higher alignment with market movements. This suggests that these models are more adept at capturing subtle nuances in news sentiment, which might be more indicative of market trends. Conversely, GPT-P5, despite processing all daily news per forex pair, showed lower correlation with both market returns and true sentiment, thus emphasizing the importance of effective prompt engineering and rigorous evaluation of prompts before deployment.

The robust correlation observed between pairs GPT-P1 and P1N through to GPT-P6 and P6N further underscores the consistency and reliability of GPT's outputs, marking them as promising candidates for real-world applications.

5.2. Potential applications in financial services

The findings of our study lend themselves to several practical applications in the financial services sector. With strategic prompt selection, the application of models like ChatGPT for financial sentiment analysis can offer actionable insights for predicting market trends. It is important to remember, though, that the choice of the ideal prompt could vary depending on the specific use case, the financial instrument under study, and the tolerance for prediction inaccuracies.

However, it is also crucial to clarify that the observed alignment between model sentiment scores and market movements does not necessarily imply predictive power over future price movements. Financial markets are complex systems influenced by a wide array of factors, including macroeconomic indicators, geopolitical events, and technical aspects. Recognizing this, a similar methodology to the one presented in this work could be applied with a more comprehensive set of input data. Instead of feeding the model with solely domain-specific news headlines, other relevant information such as economic indicators and additional market data could be included in the prompts, potentially yielding even richer insights. Therefore, while sentiment analysis using language models like ChatGPT can contribute valuable perspectives, they should be integrated into a more holistic approach to financial market analysis for optimum utility.

5.3. Limitations and future work

While our study reveals encouraging results, it also highlights some of the limitations of ChatGPT and suggests directions for future research.

Firstly, the duration of our dataset represents a limitation. The study is based on a specific time frame, and while the results are promising, the performance of the models might vary when exposed to data from different periods or under varying market conditions. Financial markets are influenced by numerous factors over time, and the short time period of our dataset might not capture all these aspects. This limitation underscores the need for further studies that span longer durations to validate and generalize our findings.

Furthermore, despite demonstrating substantial correlation with market sentiment, the models did not entirely align with the market movements, indicating that sentiment only explains a part of the variations in market prices. This observation is an important reminder of the multifaceted nature of financial markets and points towards potential avenues for future research to further enhance these models.

In the realm of generative AI, potential challenges such as model collapse, especially when heavily relying on synthetic training data, are noteworthy (Brock, Donahue, & Simonyan, 2018). Such phenomena can lead to models producing limited or identical outputs across diverse inputs (Goodfellow et al., 2014). In our study, while the use of prompts guided ChatGPT's responses, it is important to highlight that ChatGPT's foundational knowledge is based on vast and diverse real-world text data, reducing our dependency on synthetic data alone (Brown et al., 2020). This approach helps in mitigating risks associated with overfitting to synthetic patterns (Zhang, Bengio, Hardt, Recht, & Vinyals, 2021).

As the landscape of LLMs continues to evolve with remarkable speed, it is imperative that future research delves into the comparison and assessment of the unique capabilities of both existing and emerging models. For instance, newer models such as GPT-4 (OpenAI, 2023) hold potential to enhance the performance exhibited by ChatGPT in our study. Furthermore, with the availability of open-source LLMs like BLOOM, a thorough evaluation of these models, their suitability

Exhibit 1013 Page 11

for financial sentiment analysis, and their potential integration with commercial systems could prove to be a promising avenue of research. This continuous exploration of advancements in LLMs is necessary to fully leverage their potential in the dynamic and complex field of financial services.

While our initial findings highlight the capabilities of models like ChatGPT in financial sentiment analysis, it is crucial to expand our research horizon. Specifically, there is value in investigating how sentiment scores, derived from ChatGPT, influence market dynamics in the days that follow. This exploration can provide insights into the temporal relationship between sentiment and market movements.

In terms of processing time, further extensive testing under varied conditions is necessary to establish more accurate and consistent measures of response time. For operational applications that utilize LLMs services, such as the ChatGPT API, it would be advisable to have mechanisms in place to accommodate potential variations in response time due to factors like fluctuations in system load.

Taken together, the findings of our study underline the significant potential of ChatGPT in financial sentiment analysis. By pursuing the areas of future work identified here, we can move towards creating more sophisticated and accurate sentiment analysis models for financial markets with actual business value.

6. Conclusion

This work has explored the potential and adaptability of LLMs, specifically ChatGPT 3.5, in the challenging realm of financial sentiment analysis, focusing on the forex market. We employed a zero-shot prompting approach to evaluate the model's capacity to interpret and generate meaningful sentiment scores from financial text, bypassing the need for domain-specific fine-tuning. Our thorough examination was grounded in a meticulously curated and annotated dataset of forex-related news headlines, which we have made publicly available, providing a valuable resource to spur further research and innovation in the field.

Our evaluation, which extends beyond conventional metrics, revealed that prompts that accurately frame the intended task (short-term sentiment classification) show significant potential in understanding market behavior. This approach has demonstrated performance that exceeds that of FinBERT, which is an established model in the domain. Furthermore, our study underscored the critical importance of strategic prompt selection to enhance the performance and efficiency of sentiment analysis tasks, with ChatGPT demonstrating robustness and consistency across prompts.

However, while these results provide a promising outlook for the application of ChatGPT in financial sentiment analysis, they also suggest areas for further investigation. For instance, the superior understanding of overall market sentiment exhibited by prompts processing all daily news at once opens a pathway for future research exploring the integration of additional types of relevant financial data within the prompt.

In the broader landscape of AI tools in finance, our findings further substantiate the relevance and potential of LLMs. With the steady progress in this field, we anticipate more refined and robust models such as the GPT4, which could further enhance the performance of financial sentiment analysis.

While our study sheds light on the immense potential of LLMs in financial sentiment analysis, the path to real-world application is not without its challenges. Future work should look into addressing the varying time performance under different load conditions, establishing more accurate and consistent model performances.

In conclusion, our work contributes to a burgeoning field, offering a comprehensive evaluation of ChatGPT's application in financial sentiment analysis, and demonstrating its potential as a valuable tool in the realm of finance. We hope that our findings and the released dataset will serve as a springboard for further advancements in this domain. The exploration and refinement of LLMs in financial services is still in its early stages, and there is much more terrain to uncover.

CRediT authorship contribution statement

Georgios Fatouros: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **John Soldatos:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Kalliopi Kouroumali:** Data curation. **Georgios Makridis:** Writing – review. **Dimosthenis Kyriazis:** Funding acquisition, Project administration, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have included links to data/code in the manuscript.

Acknowledgments

Part of the research leading to the results presented in this paper has received funding from the European Union's funded Project FAME under grant agreement no 101092639. We would also like to express our gratitude to our FAME partners, JRC Capital Management Consultancy Research GmbH and KMcube Asset Management SA, for their invaluable contributions and expertise in the data labeling process. We extend our appreciation to the anonymous reviewers for their constructive feedback, which greatly enhanced the quality of this manuscript.

References

Arner, D. W., Barberis, J., & Buckley, R. P. (2015). The evolution of fintech: A new post-crisis paradigm. *Geological Journal of the International'l Letters*, *47*, 1271.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*, 129–151.

Bing, L. (2012). *synthesis lectures on human language technologies*, *Sentiment analysis and opinion mining*. Chicago, IL, USA: University of Illinois.

Blaskowitz, O., & Herwartz, H. (2011). On economic evaluation of directional forecasts. *International Journal of Forecasting*, *27*, 1058–1065.

Bloomberg (2023). Bloomberg media distribution. URL https://www.bloomberg.com/distribution/products/news/. accessed:May 26, 2023.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*, 1–8.

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, *27*, 1367–1403.

Chen, C., Fengler, M. R., & Härdle, Y. (2018). Textual sentiment, option characteristics, and stock return predictability.

Dakhel, A. M., Majdinasab, V., Nikanjam, A., Khomh, F., Desmarais, M. C., & Jiang, Z. M. (2023). Github copilot ai pair programmer: Asset or liability? *Journal of Systems and Software*, Article 111734.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Evans, M. D., & Lyons, R. K. (2005). Do currency markets absorb news quickly? *Journal of International Money and Finance*, *24*, 197–217.

Farimani, S. A., Jahan, M. V., Fard, A. M., & Tabbakh, S. R. K. (2022). Investigating the informativeness of technical indicators and news sentiment in financial market price prediction. *Knowledge-Based Systems*, *247*, Article 108742.

Fatouros, G., & Kouroumali, K. (2023). Forex news annotated dataset for sentiment analysis. http://dx.doi.org/10.5281/zenodo.7976208, [Data set].

Fatouros, G., Makridis, G., Kotios, D., Soldatos, J., Filippakis, M., & Kyriazis, D. (2023). Deepvar: a framework for portfolio risk assessment leveraging probabilistic deep neural networks. *Digital Finance*, *5*, 29–56.

George, A. S., & George, A. H. (2023). A review of chatgpt ai's impact on several business sectors. *Partners Universal International Innovation Journal*, *1*, 9–23.

Exhibit 1013 Page 12

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*, 1.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

JasperAI (2023). The ai in business trend report. URL https://www.jasper.ai/blog/ai-business-trend-report. accessed:May 26, 2023.

Keynes, J. M. (1937). The general theory of employment. *The Quarterly Journal of Economics*, *51*, 209–223.

Kotios, D., Makridis, G., Fatouros, G., & Kyriazis, D. (2022). Deep learning enhancing banking services: a hybrid transaction classification and cash flow prediction approach. *Journal of Big Data*, *9*, 100.

Leippold, M. (2023). Sentiment spin: Attacking financial sentiment with gpt-3. *Finance Research Letters*, Article 103957.

Liu, Z., Huang, D., Huang, K., Li, Z., & Zhao, J. (2021). Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence* (pp. 4513–4519).

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*, 35–65.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, *65*, 782–796.

Mordor Intelligence (2022). Ai in fintech market - growth, trends, covid-19 impact, and forecasts (2023–2028). URL: https://www.mordorintelligence.com/industry-reports/ai-in-fintech-market. Accessed:April 20, 2023.

OpenAI (2023). Gpt-4 technical report. arXiv:2303.08774.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, Volume 1 (long papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N18-1202, URL https://aclanthology.org/N18-1202.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125.

Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, *108*, 42–49.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*, 9.

Refaeli, D., & Hajek, P. (2021). Detecting fake online reviews using fine-tuned bert. *7*, In *Proceedings of the 2021 5th International Conference on E-Business and Internet* (pp. 6–80).

Sallam, M. (2023). Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare* (p. 887). MDPI.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, D., Castagné, A. S., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transctions on Information Systems (TOIS)*, *27*, 1–19.

Siering, M., Muntermann, J., & Rajagopalan, B. (2018). Explaining and predicting online review helpfulness: The role of content and reviewer-related signals. *Decision Support Systems*, *108*, 1–12.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*, 1139–1168.

Thoppilan, R., Freitas, D. De., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., et al. (2022). Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45).

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., et al. (2023). Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564.

Yeshayahou, K. (2021). Israeli co similarweb files for nyse ipo. URL: https://en.globes.co.il/en/article-israeli-co-similarweb-files-for-nyse-ipo-1001367823. Accessed:May 11, 2023.

Yue, T., Au, D., Au, C. C., & Iu, K. Y. (2023). Democratizing financial knowledge with chatgpt by openai: Unleashing the power of technology. Available at SSRN 4346152.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*, 107–115.