IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 July 2024, accepted 9 August 2024, date of publication 19 August 2024, date of current version 27 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3445413

RESEARCH ARTICLE

Large Language Models and Sentiment Analysis in Financial Markets: A Review, Datasets, and Case Study

CHENGHAO LIU[1], ARUNKUMAR ARULAPPAN[2], (Member, IEEE), RANESH NAHA[3], (Member, IEEE), ANIKET MAHANTI[1,4], (Senior Member, IEEE), JOARDER KAMRUZZAMAN[5], (Senior Member, IEEE), AND IN-HO RA[6], (Member, IEEE)

[1]School of Computer Science, The University of Auckland, Auckland 1010, New Zealand
[2]School of Computer Science Engineering and Information Systems, VIT University, Vellore 632014, India
[3]School of Information Systems, Queensland University of Technology, Brisbane, QLD 4000, Australia
[4]Department of Computer Science, University of New Brunswick, Saint John, NB E2K 5E2, Canada
[5]Centre for Smart Analytics, Federation University Australia, Melbourne, VIC 3806, Australia
[6]School of Software, Kunsan National University, Gunsan 54150, South Korea

Corresponding author: Arunkumar Arulappan (arunkumar.a@vit.ac.in)

This work was supported by the School of Computer Science Engineering and Information Systems, Vellore Institute of Technology.

ABSTRACT This paper comprehensively examines Large Language Models (LLMs) in sentiment analysis, specifically focusing on financial markets and exploring the correlation between news sentiment and Bitcoin prices. We systematically categorize various LLMs used in financial sentiment analysis, highlighting their unique applications and features. We also investigate the methodologies for effective data collection and categorization, underscoring the need for diverse and comprehensive datasets. Our research features a case study investigating the correlation between news sentiment and Bitcoin prices, utilizing advanced sentiment analysis and financial analysis methods to demonstrate the practical application of LLMs. The findings reveal a modest but discernible correlation between news sentiment and Bitcoin price fluctuations, with historical news patterns showing a more substantial impact on Bitcoin's longer-term price than immediate news events. This highlights LLMs' potential in market trend prediction and informed investment decision-making.

INDEX TERMS Large language model, Bitcoin price, sentiment analysis, machine learning, market dynamics.

I. INTRODUCTION

Sentiment analysis (SA) in financial markets has emerged as a critical study area, particularly given its widespread application in specific sectors like the stock market [1], [2], [3], [4]. This analytic approach primarily aims to discern individuals' attitudes, evaluations, and opinions regarding various entities and products. In this context, behavioral economics becomes pertinent as it delves into the psychological aspects of investor behaviors, considering the influence of social, cultural, and emotional factors on decision-making processes [5]. These factors often play a significant role in explaining market anomalies [6].

Furthermore, the sentiment expressed in news, especially those covering political, social, economic, or emotional events disseminated through social media, profoundly influences investor behavior [7], [8]. As a result, information sourced from online newsgroups, social networks, and stock discussion forums has become increasingly valuable for informed business decision-making. Recently, a significant amount of research has been carried out by fundamentally analyzing unstructured text data through machine learning involving supervised and unsupervised learning methods. LLMs emerged due to large-scale data and increased

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

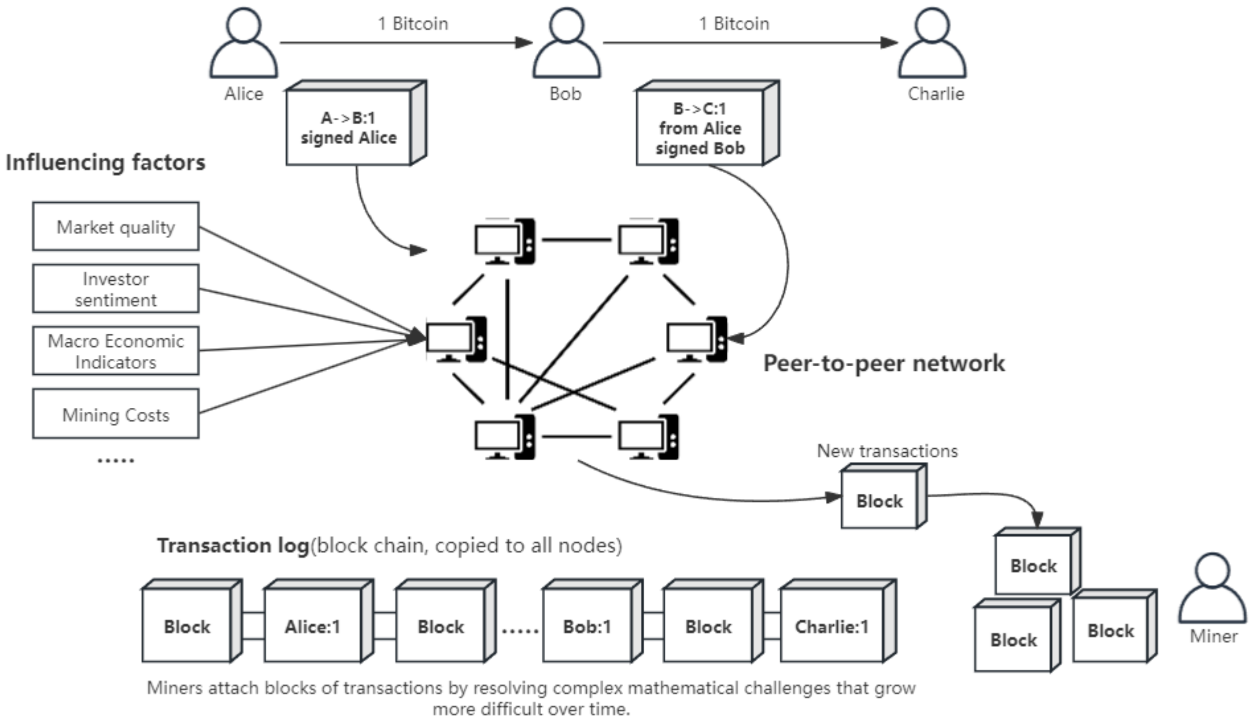

FIGURE 1. **Bitcoin's approach to transaction flow and validation.**

computational power available [9]. Armed with a wide range and variety of training data, these models have shown remarkable proficiency in mimicking human language skills, resulting in significant transformations across various fields, including the financial domain [10]. Applying LLMs to sentiment analysis represents an innovative shift, where the traditional sentiment analysis challenges are reinterpreted and addressed through more advanced computational approaches [11]. Their effectiveness is particularly notable in tasks that require deep contextual understanding and nuanced language interpretation, such as predicting market trends, analyzing investor sentiments, and interpreting financial news [10], [12], [13]

Despite the growing interest in using LLMs for sentiment analysis, especially in financial markets, there remains a significant gap in understanding the extent and nature of their impact on financial instruments, particularly cryptocurrencies like Bitcoin. Existing literature predominantly focuses on the technical capabilities of LLMs without adequately exploring their practical implications in financial sentiment analysis. Our study seeks to bridge this gap by not only categorizing various LLMs and their applications in financial markets but also by empirically investigating the correlation between news sentiment, as processed by these models, and Bitcoin price movements. This approach aims to provide a more nuanced understanding of the role of media sentiment in cryptocurrency markets. To achieve this, our study will answer the following research questions **(RQ)**:

1) **RQ1**: How does the classification, data collection, and application of LLMs in sentiment analysis influence their effectiveness in financial markets?
2) **RQ2**: What is the correlation between news sentiment, as analyzed by LLMs, and the price of cryptocurrencies like Bitcoin?

This paper makes a significant contribution to the field of financial sentiment analysis by integrating the advanced capabilities of LLMs with the dynamic realm of Bitcoin and cryptocurrency markets. The study stands out for its comprehensive examination of various LLMs, including BERT, FinBERT, and ChatGPT, within the specific context of financial sentiment analyses [12], [14], [15]. This area is particularly challenging due to the nuanced language and investor sentiments intrinsic to market dynamics [16]. The systematic categorization and analysis of these LLMs in the paper illuminate their individual strengths and collective potential in enhancing financial market analytics. The focus on the unique features and applications of these LLMs in the financial domain reveals new insights into their transformative role in market trend prediction and investment decision-making.

By identifying a modest but discernible correlation between news sentiment and Bitcoin prices, the paper contributes valuable empirical evidence to the understanding of cryptocurrency market dynamics. This insight is crucial for a range of stakeholders, including investors, financial analysts, and policymakers, who navigate the complexities

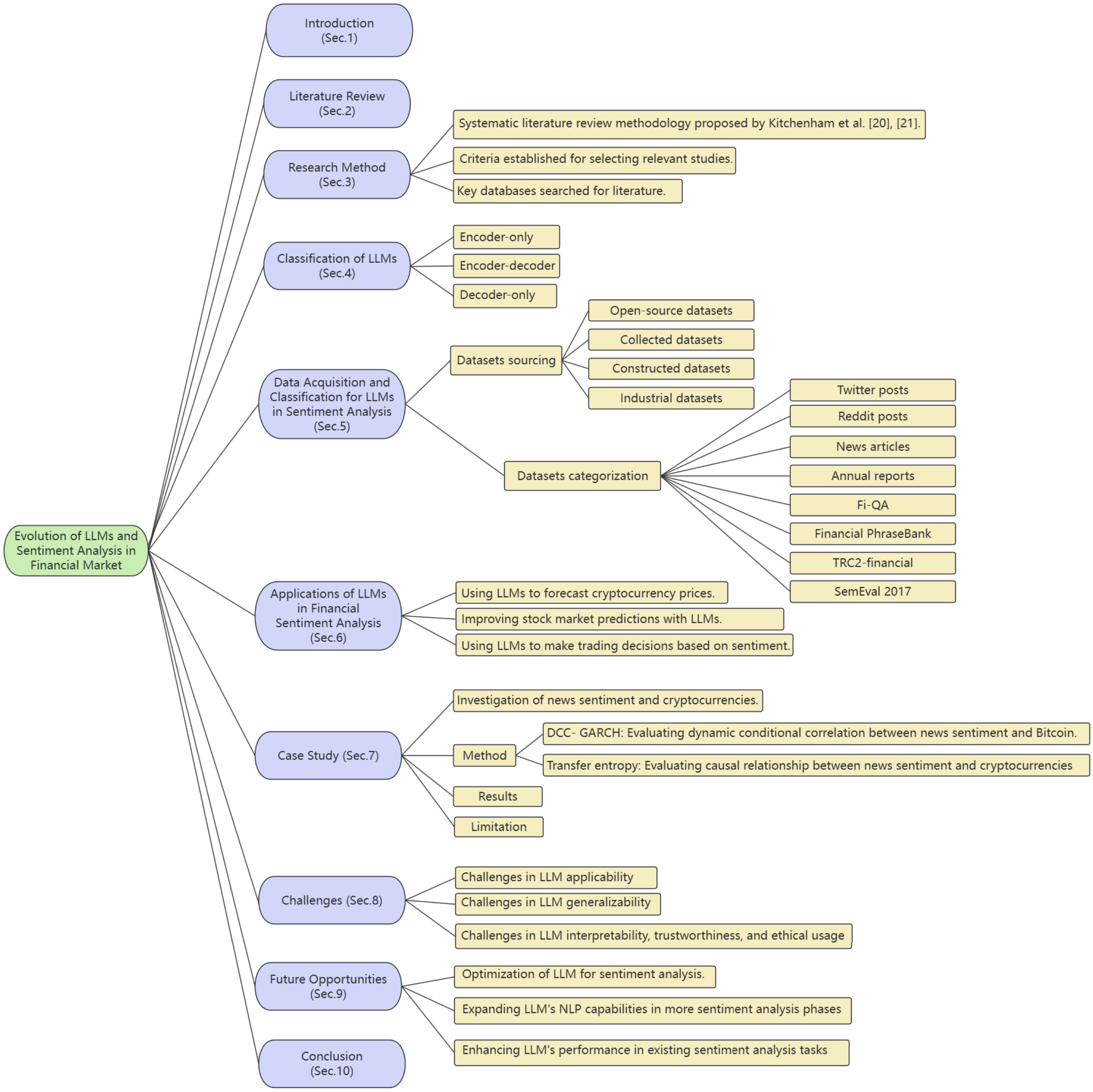
Evolution of LLMs and Sentiment Analysis in Financial Market
Introduction (Sec.1)
Literature Review (Sec.2)
Research Method (Sec.3)
Systematic literature review methodology proposed by Kitchenham et al. [20], [21].
Criteria established for selecting relevant studies.
Key databases searched for literature.
Classification of LLMs (Sec.4)
Encoder-only
Encoder-decoder
Decoder-only
Data Acquisition and Classification for LLMs in Sentiment Analysis (Sec.5)
Datasets sourcing
Open-source datasets
Collected datasets
Constructed datasets
Industrial datasets
Datasets categorization
Twitter posts
Reddit posts
News articles
Annual reports
Fi-QA
Financial PhraseBank
TRC2-financial
SemEval 2017
Applications of LLMs in Financial Sentiment Analysis (Sec.6)
Using LLMs to forecast cryptocurrency prices.
Improving stock market predictions with LLMs.
Using LLMs to make trading decisions based on sentiment.
Case Study (Sec.7)
Investigation of news sentiment and cryptocurrencies.
Method
DCC- GARCH: Evaluating dynamic conditional correlation between news sentiment and Bitcoin.
Transfer entropy: Evaluating causal relationship between news sentiment and cryptocurrencies
Results
Limitation
Challenges (Sec.8)
Challenges in LLM applicability
Challenges in LLM generalizability
Challenges in LLM interpretability, trustworthiness, and ethical usage
Future Opportunities (Sec.9)
Optimization of LLM for sentiment analysis.
Expanding LLM's NLP capabilities in more sentiment analysis phases
Enhancing LLM's performance in existing sentiment analysis tasks
Conclusion (Sec.10)

FIGURE 2. Structure of this paper.

of these emerging markets. Additionally, the discussion of challenges and future directions for LLMs in sentiment analysis highlights both the current capabilities and potential growth areas for these models in financial applications. Fig. 1. illustrates the process of Bitcoin transactions within a peer-to-peer network, showcasing how transactions are signed and added to the blockchain through mining, while also highlighting various factors that can influence Bitcoin's market price.

This study is structured as follows: Section II presents the existing literature closely related to sentiment analysis in financial markets. Section III outlines our research methodology. Sections IV to VI are collectively focused on addressing RQ1. Section IV presents a detailed classification of LLMs in sentiment analysis. Section V discusses the data collection method and categorization. Section VI explores the applications of LLMs in sentiment analysis. Section VII provides a case study aimed at answering our RQ2, offering practical insights into applying these models in a real-world scenario. Sections VIII and IX discuss the challenges that should be overcome when employing LLMs to solve sentiment analysis tasks and highlight promising opportunities and directions for future research. The conclusions of our study are presented in section X. The overall organization of the paper is presented in Fig. 2.

II. LITERATURE REVIEW

Among the various LLMs, BERT (Bidirectional Encoder Representations from Transformers) [14] has set a new precedent in natural language processing by understanding the context of a word in a sentence more holistically. BERT's architecture has been utilized in the financial sector to create FinBERT [12], a model specifically fine-tuned to grasp the subtleties of financial jargon and sentiment. FinBERT [12] excels in interpreting complex financial reports, earnings calls, and market analysis, providing more accurate sentiment predictions than general-purpose models [17]. Additionally, Ploutos [18], another financial LLM, demonstrates superior performance in predicting stock movements. This model uniquely integrates textual and numerical data using a mixture of experts architecture, enhancing its ability to deliver precise explanations for its predictions. A further groundbreaking LLM is ChatGPT [15], which has been instrumental in enhancing interactive financial analysis. ChatGPT's ability to engage in human-like conversations and provide detailed, contextually relevant responses has been utilized in customer service automation, financial advisory, and real-time market analysis [19]. This model's sophisticated understanding of queries and ability to generate coherent and context-aware responses make it an invaluable tool in the dynamic world of finance.

Few recent studies focus on various applications and advancements of LLMs in financial sentiment analysis. In a recent study, Sharma et al. [20] explored the use of generative models like ChatGPT for sentiment analysis. These models enhance sentiment analysis by augmenting datasets with synthetic labeled data and simulating human sentiment expression, particularly for tasks like sarcasm detection. Key challenges include maintaining the quality and consistency of generated data and addressing inherent biases. By overcoming these issues, the potential for sentiment analysis in real-world applications can be significantly enhanced. The architectures and applications of large language models, including their use in sentiment analysis presented by Raiaan et al. [21]. They categorize different LLMs, such as GPT-3, and explore their applications in various domains, including finance. The paper also addresses the challenges and open issues in deploying LLMs for sentiment analysis, such as data scarcity and model interpretability. The increasing role of generative AI models, such as GPT-3, in business and finance is discussed in [22]. The work highlights the potential of these models to generate realistic financial data, perform sentiment analysis, and support decision-making processes. The paper also explores the ethical implications and regulatory challenges associated with using generative AI in financial markets.

Dong et al. [23] investigate the application of LLMs for extracting relevant information from financial documents. The authors employ GPT-3 to analyze annual reports, earnings call transcripts, and other financial texts to identify key sentiment indicators and predict stock price movements. The study shows that LLMs can effectively process and interpret large volumes of text data, providing valuable insights for investors and analysts. Farimani et al. [24] investigate the efficiency and accuracy of using LLMs like GPT-3 for sentiment analysis in the financial market. The authors compare the performance of LLMs with traditional models, demonstrating significant improvements in capturing nuanced sentiments and predicting market trends based on financial news and social media data. Another early review [25] emphasizes the potential of both BERT and GPT-2 in advancing financial sentiment analysis through improved feature mapping techniques, leveraging their respective strengths in understanding context and generating relevant text.

While previous studies have significantly advanced financial sentiment analysis using models like FinBERT and integrated approaches combining sentiment indices with predictive models, our approach introduces a novel perspective by leveraging more specific aspects like classification, data collection and application with a case study. A central element of this research is the empirical investigation into the correlation between news sentiment, as analyzed by LLMs, and Bitcoin price movements. This case study is particularly relevant given the growing influence of cryptocurrencies like Bitcoin in global financial markets. Bitcoin serves as a benchmark for the digital currency landscape, characterized by its volatility, decentralized nature, and sensitivity to public sentiment and news [26]. The study addresses the pressing need to understand Bitcoin's market behavior due to its escalating impact on retail and institutional investors and its potential in reshaping financial technology and monetary transactions [27].

III. RESEARCH METHOD

This literature review adheres to the methodology proposed by Kitchenham et al. [28], [29]. Following the guidelines provided by Kitchenham et al. [28], our methods included two main steps: planning and conducting the review. Established academic databases were utilized to gather the relevant literature, including *Web of Science, IEEE Xplore, Springer, arXiv, and UoA(University of Auckland) Library*. The following sections describe the methodology used to source and evaluate the chosen literature. Specifically, Fig. 3 presents the structure of the literature review.

Our manual search encompasses four critical databases known for their comprehensive collection of scientific papers. The methodology involved a multi-step process, beginning with creating a keyword dictionary instrumental in the initial search across these databases. Our search string should

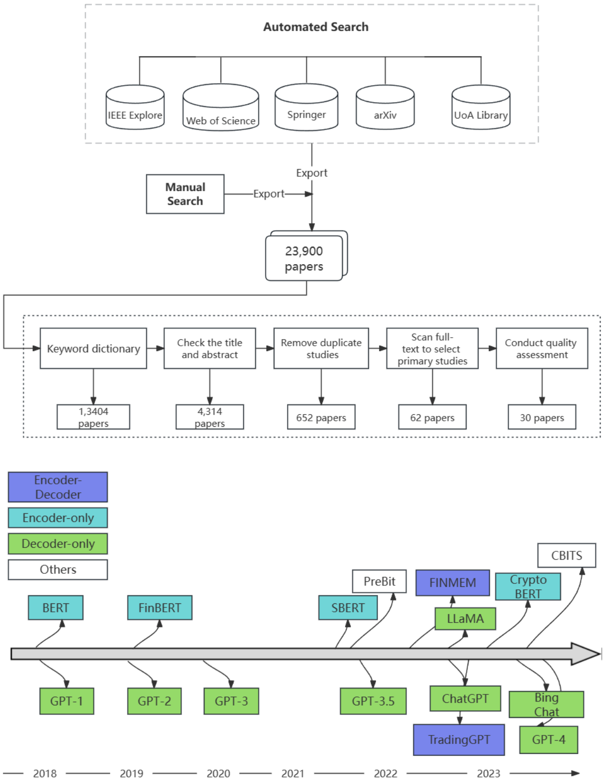

FIGURE 3. LLMs classification and literature review methodology in financial sentiment analysis.

combine two sets of keywords: one related to sentiment analysis and the other to LLMs. If the paper contains both types of keywords, it is more likely that it is the paper we need. The complete set of search keywords is as follows:

1) *Keywords related to sentiment analysis: Sentiment detection, Opinion mining, Emotional analytics, Affective computing, Polarity classification, Subjectivity analysis, Sentiment scoring, Mood analysis, Opinion polarity, Sentiment quantification, Emotion recognition, Tone analysis, Sentiment lexicons, Sentiment metrics, Textual affect detection, Semantic orientation, Sentiment strength, Sentiment benchmarks, Sentimental analysis tools, Review analysis, Consumer sentiment, Investor sentiment, Market sentiment, Brand sentiment, Social sentiment, Sentiment correlation, Aspect-based sentiment analysis, Sentiment summarization, Sentimental classification, Sentimental interpretation.*
2) *Keywords related to LLMs: LLM, Language Model, Large Language Model, Pre-trained, PLM, Pre-training, NLP, Natural Language Processing, DL, Deep Learning, ML, Machine Learning, ChatGPT, Neural Network, Transfer Learning, Sequence Model, T5, GPT, Codex, BERT, Transformer, Attention Model, AI, Artificial Intelligence.*

We included keywords like Machine Learning and Deep Learning, alongside other terms not directly related to Large Language Models (LLMs), in our search criteria. This broader approach aims to ensure we don not overlook any relevant research, thereby expanding our search scope during automated searches.

Upon retrieving these papers, the next step involved a detailed examination of the titles and abstracts, which allowed to determine the relevance of each paper to the research objectives based on inclusion and exclusion criteria. We designed these criteria following several several state-of-the-art papers [30], [31], as shown in Table 1, so that the selected documents can directly address our topic. We drop duplicated studies across multiple databases to refine our dataset, streamlining our literature collection.

TABLE 1. Inclusion criteria and exclusion criteria.

Inclusion criteria

1) It is claimed in the paper that LLM is used.
2) Statements within the paper about conducting sentiment analysis.
3) Papers where the complete text is available.

Exclusion criteria

1) Repetitive papers or similar research by the same authors in different forms.
2) Non-English written literature.
3) Studies belonging to keynotes, books, panels, monographs or venues need to execute a complete peer-review process.
4) Papers that reference using LLMs but do not detail the techniques used.

Following the curation of unique studies, we proceeded to a more in-depth review, scanning the full text of each selected paper. A thorough quality assessment can help mitigate biases that may arise from low-quality studies and guide readers on where to approach conclusions with caution [32]. We developed a set of ten Quality Assessment Criteria (QAC), detailed in Table 2. These criteria evaluate the papers' relevance, clarity, validity, and importance. The final stage in our search process was to conduct a quality assessment of these primary studies, evaluating them against predefined criteria to ensure that only the most rigorous and relevant research was included in our analysis.

The systematic literature review on LLMs for sentiment analysis acknowledges the risk of missing key studies due to potential gaps in keyword summarization. To mitigate this, a dual approach combining manual review and automated searches was utilized, with keywords derived from authoritative sources and forward and backward snowballing techniques employed to ensure thoroughness. Additionally, to counter study selection bias, defined inclusion and exclusion criteria were established, and a QAC framework was implemented, with ambiguous cases receiving manual scrutiny. This blend of strategies aimed to balance efficiency with meticulousness, reducing biases and enhancing the review's validity.

TABLE 2. Checklist of quality assessment criteria (QAC) for studies on LLMs in sentiment analysis.

Quality Assessment Criteria
1) Is the research focused on tasks related to sentiment analysis?
2) Does the research incorporate the use of LLMs?
3) Does the research represent an original primary rather than a secondary study like a systematic literature review or survey?
4) Has the research been published in a well-regarded academic or professional venue?
5) Is the purpose or rationale of the research clearly stated?
6) Does the study provide a detailed explanation of the methodologies used?
7) Is there a detailed explanation of the experimental setup, including conditions and data details?
8) Is there conclusive verification of the results obtained in the study?
9) Does the research paper address its key findings and acknowledge any limitations?
10) Does the study significantly contribute to the academic or industrial fields?

IV. CLASSIFICATION OF LARGE LANGUAGE MODELS IN SENTIMENT ANALYSIS

This section explores the classification of LLMs in the context of sentiment analysis, emphasizing how their size and architecture impact their effectiveness. We categorize LLMs based on their structural design, distinguishing between encoder-only, encoder-decoder, and decoder-only models, each with distinct capabilities in processing natural language. For an in-depth exploration of the relevant literature, we have included a comprehensive summary in Table 4.

A. LARGE LANGUAGE MODELS

Pre-trained language models (PLMs) have proven highly effective in various natural language processing (NLP) tasks, as evidenced in several studies [33], [34]. Researchers have noted that increasing the size of these models significantly boosts their capabilities, particularly when the size of parameters exceeds a certain point [35], [36]. The designation "Large Language Model" (LLM) is used to differentiate language models based on their size, primarily referring to PLMs with a larger scale [37]. However, it is important to mention that there is no widely agreed-upon standard in the literature for the minimum parameter size for an LLM, as its efficiency is linked to the dataset's size and the total computing power used. In our study, we follow the classification and taxonomy of LLMs introduced by Pan et al. [38], dividing mainstream LLMs into three categories based on their architecture: encoder-only, encoder-decoder, and decoder-only. This classification and the corresponding models are depicted in Fig. 3.

1) ENCODER-ONLY LLMs

Encoder-only LLMs are a specific type of neural network framework that employs solely the encoder part of the model. The primary role of the encoder is to process and transform the input text into a hidden representation. This representation is critical in understanding the connections among words and the general context of the sentence. Prominent examples of encoder-only LLMs include BERT [14] and various adaptations of it [12], [39], [40]. BERT, in particular, is built on the encoder architecture of the Transformer [41]. Its unique feature is the bidirectional attention mechanism, which allows it to analyze the context to the left and right of each word concurrently during its training phase. In the financial domain, other prominent models like FinBERT [12], CryptoBERT [42], and SBERT [26]have been widely employed.

These models distinguish themselves from the original BERT [14] by enhancing the architecture to include novel pre-training tasks or adjusting to different data modalities, thereby improving their effectiveness for finance-related tasks. For instance, FinBERT [12] is an adaptation of BERT [14] that is pre-trained explicitly on financial corpora and fine-tuned to perform sentiment analysis within the financial domain, achieving an accuracy of 0.86 and an F1-Score of 0.84. Similarly, CryptoBERT [42], which is also grounded in the BERT [14] model, undergoes fine-tuning on a cryptocurrency-specific corpus, yielding heightened accuracy in the sentiment classification of texts related to cryptocurrencies. It achieved accuracy scores of 55.60 and an F1-Score of 55.79 among five models for the StockTwits[1] data, which contains 1.875 million posts. These models have demonstrated their proficiency in various applications, such as predicting market movements, analyzing investor sentiments, and automating financial report summaries, showcasing their transformative impact on the financial analytics landscape.

2) ENCODER-DECODER LLMs

Encoder-decoder LLMs integrate both the encoder and decoder components [41]. The encoder component converts the input text into a hidden representation, adeptly grasping the fundamental structure and meaning. This confidential representation is a transitional language, facilitating the connection between input and output formats. On the other hand, the decoder leverages this hidden representation to produce the desired output text, transforming the abstract representation into specific, contextually appropriate phrases. Within this context, the memory module of models like FINMEM [43] stands out. It mirrors human cognitive processes, providing clear interpretability and flexibility for real-time adjustments. This feature enhances the model's utility in financial trading by allowing it to hold on to essential information for extended periods, which is crucial for complex decision-making. FINMEM outperformed in trading five different stocks, achieving the highest Sharpe ratio of 2.6789 and lowest max drawdown of 10.7996%. Another example is TradingGPT [44], an innovative LLM multi-agent framework endowed with layered memories. The ability of TradingGPT to navigate through financial data and

[1] https://stocktwits.com/

its application in trading exemplifies how encoder-decoder LLMs can be potent tools in enhancing trading strategies [44].

3) DECODER-ONLY LLMs

Decoder-only LLMs exclusively use the decoder module to produce the intended output text. They follow a unique training approach focusing on sequential prediction [45]. Contrary to the encoder-decoder framework, where the encoder handles the input text, the decoder-only structure starts from a base state and sequentially predicts tokens, thereby progressively constructing the output text. This method heavily depends on the model's proficiency in grasping and predicting language structure, syntax, and context. Key examples of this architecture include the GPT series models such as GPT-1, GPT-2, GPT-3, GPT-4, and their significant variant, ChatGPT.[2] [45], [46], [47], [48]. The GPT series has shown promising performance in financial sentiment analysis, not only for Twitter news but also in terms of accuracy, recall, and F1-score across different forex pair news [49], [50]. These models demonstrate their capability to excel in financial contexts, highlighting their potential for improving sentiment analysis and market prediction tasks.

These models can execute downstream tasks with minimal input, often requiring just a handful of examples or straightforward instructions. This attribute eliminates the need for additional prediction heads or extensive fine-tuning processes, rendering them particularly valuable in sentiment analysis research. For instance, recent developments in the industry have witnessed Google unveiling Bard. At the same time, Meta has introduced its models, LLaMA [51] and LLaMA2 [52], alongside Microsoft's foray with Bing Chat.[3] One application of LLaMA in the realm of financial sentiment analysis is demonstrated by FinMA, a version of LLaMA specifically fine-tuned for this task, which recorded the highest F1-score of 0.87 on the FiQA dataset [53]. Furthermore, LLaMA2 has proven effective, reaching an accuracy of 84.03% through supervised learning and aligning financial texts [54]. These developments highlight the capabilities of LLaMA models in sentiment analysis, particularly their proficiency in the precise interpretation and assessment of financial sentiments.

V. DATA ACQUISITION AND CLASSIFICATION FOR LLMs IN SENTIMENT ANALYSIS

This section examines the methodologies employed in collecting and utilizing datasets for sentiment analysis in LLMs. This section underscores the pivotal role of data in training LLMs, emphasizing the need for diversity and comprehensiveness in dataset collection to enhance model performance in varied contexts [55]. We explore the systematic process of dataset categorization, preprocessing, and formatting, which is essential for aligning data with the model's training objectives and processing needs.

A. SOURCING DATASETS FOR TRAINING LARGE LANGUAGE MODELS

Data is a vital and essential component in training Large Language Models (LLMs), significantly influencing their generalization capabilities, efficiency, and overall performance [55]. An ample amount of high-quality and varied data enables models to thoroughly learn features and patterns, fine-tune their parameters, and maintain dependability during validation and testing.

Our initial focus is on examining the methodologies for dataset acquisition. Through this analysis of data collection techniques, we have categorized the sources of data into four groups: open-source datasets, datasets that are actively collected, datasets that are specifically constructed, and datasets derived from industrial sources. *Open-source datasets* [56], [57] are publicly available data compilations typically distributed via open-source platforms or repositories. An example of this is the FiQA [56] dataset, a substantial new dataset featuring Question-Answering pairs focused on financial reports crafted by experts in finance. The credibility of these datasets is bolstered by their open-source status, enabling community-based updates and ensuring their reliability for scholarly research.

The Financial PhraseBank, first introduced by Malo et al. [58], consists of 4,845 English sentences randomly selected from financial news articles in the LexisNexis database. These sentences were annotated by 16 experts in finance and business who evaluated how the information could influence the stock prices of the companies discussed. Furthermore, the dataset includes information about the level of agreement among the annotators regarding the sentiments expressed in the sentences.

TRC2-financial is a specialized subset of the TRC24[4] collection from Reuters, which encompasses 1.8 million news articles released between 2008 and 2010. This subset specifically contains 46,143 documents, totaling nearly 29 million words and close to 400,000 sentences [12].

SemEval 2017 Task 5 focuses on fine-grained sentiment analysis (FSA) of news headlines and microblogs [59]. The training set for this task includes 1,142 financial news headlines and 1,694 microblog posts, each annotated with target entities and their corresponding sentiment scores. The test set comprises 491 financial news headlines and 794 posts [11].

Collected datasets [26], [60] are compiled by researchers from diverse sources, such as significant websites, forums, blogs, and social media. Researchers often extract data from sources like Twitter and Reddit for datasets specifically tailored to their research inquiries.

Constructed datasets [12] are researcher-generated datasets derived from modifying or enhancing collected datasets to align with specific research goals. Manual or semi-automatic modifications can include creating domain-specific tests, annotated datasets, or synthetic data.

[2] https://chat.openai.com/

[3] https://www.microsoft.com/en-us/edge/features/bing-chat

[4] https://trec.nist.gov/data/reuters/reuters.html

TABLE 3. Data types of datasets involved in prior studies.

Data Type	Number of Studies	References	Number of Data Points	Data Range	Data Source
Twitter Post	4	Zou and Herremans [61]	9,435,437	2015 to 2021	Collected datasets
		Raheman et al. [62]	100,000	July to December of 2021	Constructed datasets
		Nguyen et al. [27]	9,198	2021 to 2022	Collected datasets
		Kulakowski and Frasincar [42]	496,000	11 to 24 July, 2018	Collected datasets
Reddit Post	2	Ortu et al. [63]	33,000	January 2017 to January 2021	Collected datasets
		Kulakowski and Frasincar [42]	172,000	1 May 2021 to 30 April 2022	Collected datasets
News Article	6	Bashchenko [26]	33,283	February 2015 to June 2021	Collected datasets
		Kulakowski and Frasincar [42]	1.875 million	1 November 2021 to 30 June 2022	Collected datasets
		Kim et al. [60]	18,005	19 January 2018 to 16 April 2022	Collected datasets
		Yu et al. [43]	15,500	15 August 2021 to 25 April 2023	Constructed datasets
		Li et al. [44]	-	15 August 2020 to 15 August 2023	Collected datasets
		Fazlija and Harder [64]	447,279	3 January 2007 and 26 November 2013	Open-source datasets
Annual Report	1	Gupta [65]	24,200	2002 to 2023	Open-source datasets
Fi-QA	2	Deng et al. [66]	8,281	1999 to 2019	Open-source datasets
		Wu et al. [10]	8,281	1999 to 2019	Open-source and Industrial datasets
Financial PhraseBank	2	Araci [12]	4,846	-	Open-source datasets
		Fazlija and Harder [64]	4,846	-	Open-source datasets
TRC2-financial	1	Araci [12]	46,143	2008 to 2010	Open-source datasets
SemEval 2017	1	Mishev et al. [11]	2150	August and November 2015	Open-source datasets

Industrial datasets [10] sourced from commercial or industrial entities contain proprietary data and are essential for research addressing real-world business contexts.

Acknowledging that certain studies utilize diverse datasets encompassing various categories is essential. For instance, Wu et al. [10] trained BloombergGPT using multiple datasets, e.g., complex table datasets and question-answering pairs.

B. VARIETY OF DATASETS IN EXISTING LLMs FOR SENTIMENT ANALYSIS STUDIES

The data types are crucial in determining the architecture and choice of LLMs, as they directly affect the extraction of implicit features and the decisions made by the model [67]. The selection of specific data types can significantly influence the LLMs' overall effectiveness and ability to generalize. In our research, we explore and categorize the various types of financial datasets used in studies of LLMs for sentiment analysis. By examining how data types relate to model architectures and their performance, we aim to highlight the importance of data types in the effectiveness of LLMs for sentiment analysis.

We classified the data types of all datasets into five categories: Twitter posts, Reddit posts, News articles, Annual reports, and Fi-QA. Table 3 describes the specific data included in the data types corresponding to the datasets we summarized from the 15 studies.

VI. APPLICATIONS OF LLMs IN FINANCIAL SENTIMENT ANALYSIS

This section delves into LLMs' diverse and transformative applications in financial sentiment analysis. In recent years, integrating advanced LLMs into the financial sector has marked a significant evolution in how financial data, market trends, and investor sentiments are analyzed and interpreted. This section explores how LLMs predict market trends, optimize trading strategies, and forecast stock prices.

A. PREDICTIVE ANALYTICS IN CRYPTOCURRENCY MARKETS USING LLMs

This section explores the application of LLMs for predicting cryptocurrency market trends, with a particular focus on integrating sentiment analysis into these predictions. The potential of LLMs to distill sentiment from vast datasets offers a novel dimension to the forecasting models, as evidenced by several recent studies. Zou and Herremans [61] introduced a pioneering multimodal model, PreBit, specifically designed to anticipate significant Bitcoin price movements. Bashchenko [26] provided insights that counter the notion of Bitcoin's value being purely speculative, demonstrating that non-endogenous news carries fundamental information affecting Bitcoin prices.

Raheman et al. [62] highlighted the practical advantages of interpretable AI and NLP methods over non-explainable alternatives, suggesting that transparency in AI could lead to more valuable applications in the financial sector. Ider and Lessmann [68] demonstrated the advantages of refining FinBERT with weakly labeled data, illustrating how even imprecisely labeled datasets can significantly improve text-based feature prediction and forecasting accuracy for cryptocurrency returns. Their study utilized a dataset comprising 433 test samples, with a noteworthy agreement rate of 92.6% among all 16 expert labels. This approach facilitated the development of predictive models for Bitcoin and Ethereum that substantially outperformed baseline models, achieving gains of 0.572 and 0.501, respectively.This evidence underscores the efficacy of leveraging weak labels in enhancing the performance of financial prediction models, particularly in the volatile domain of cryptocurrency markets.

Ortu et al. [63] investigated cryptocurrency price prediction by analyzing social sentiment data from GitHub and Reddit, employing a pre-trained BERT-based model to synthesize emotional and sentiment indicators from social media commentary into hourly and daily series datasets. Their findings indicated that incorporating these social sentiment metrics markedly enhances the predictive accuracy for the daily pricing of Bitcoin and Ethereum. The research highlights a significant inverse relationship between negative sentiment and price volatility within the Bitcoin market, suggesting that users might interpret volatility as a speculative opportunity. In contrast, the Ethereum market sentiment is predominantly influenced by emotional arousal, which shows a substantial positive correlation with negative sentiment, indicating that community reactions are more emotionally driven rather than directly related to price movements.

Building on these findings, Nguyen et al. [27] explored the distinctive impact of ChatGPT-based sentiment indicators on Bitcoin returns, revealing its adeptness at sentiment detection. The study's results prompt further investigation into how Generative AI might enhance financial data analysis and social media sentiment interpretation, potentially unlocking more sophisticated market insights. This research opens up new pathways for sentiment analysis in financial markets, leveraging AI technologies.

B. SENTIMENT-DRIVEN LLM STRATEGIES FOR FINANCIAL TRADING

Developing a robust trading strategy is crucial in the volatile realm of financial markets, where integrating sentiment analysis and LLMs can provide a competitive edge. Kim et al. [60] leveraged an LLM adapted to the crypto domain to parse crypto news sentiments called CBITS. Their research demonstrates that trading strategies augmented with sentiment scores significantly outperform conventional models, underscoring the efficacy of sentiment-based trading approaches. Backtesting various Bitcoin trading strategies, their study reveals that models employing TabNet combined with RoBERTa, specifically the TabNet RoBERTa top 10, yield the highest profit, recording an impressive gain of 304.65%. In contrast, other models assessed during the same test period generated negative returns.

Yu et al. [43] introduced FINMEM, an innovative LLM-based framework crafted for financial decision-making. This framework is structured around three central modules: Profiling, which tailors the agent to specific investor profiles; Memory, which processes financial information in a layered manner akin to human cognitive structures, facilitating deeper assimilation of financial data; and Decision-making, which translates the processed information into actionable investment strategies. The adaptability of FINMEM, particularly its memory module, provides a level of interpretability that mirrors human trading logic, coupled with the capability for real-time adjustment to optimize trading decisions.

Li et al. [44] took the concept further by developing an LLM multi-agent framework with layered memories called TradingGPT. The LLMs at the heart of this framework act as decision-making cores for trading agents, utilizing the layered memory system to synthesize historical data and current market conditions. This innovative approach enables the agents to engage in strategic dialogues with peers, refine their investment choices, and uphold a diverse yet robust decision-making process informed by their unique trading personas.

Curtó et al. [69] provided empirical evidence showcasing the adaptability of LLM-informed strategies to the dynamic bandit problem, a standard paradigm in trading strategy formulation. Their experiments underscore the ability of LLMs to navigate the complexities of the financial markets, yielding a strategy that competes favorably with traditional methods even in unpredictable scenarios.

Gupta [65] aimed to streamline the analysis of Annual Reports across various firms by harnessing the analytical prowess of LLMs. A machine learning model was trained using these insights as predictive features by distilling insights from the LLMs into a quantitatively styled dataset and supplementing it with historical stock prices. The walk-forward testing indicated that such a model could significantly outperform benchmarks like the S&P 500 returns, underscoring the potential of GPT3.5 to revolutionize trading strategies. The research revealed that the model, when used to select the top k stocks, consistently generated higher returns than the S&P 500. Notably, the returns were inversely related to the value of k, with lower k values correlating with higher returns. This outcome indicates that the stocks predicted as top performers by the GPT model indeed yielded better financial results.

C. ENHANCING STOCK MARKET FORECASTING WITH LLMS

Our analysis underscores the broad utility of LLMs in stock price prediction through sentiment analysis, showcasing their versatility across various financial applications. Araci introduced FinBERT, a model tailored for the financial sector, demonstrating superior capabilities in economic text mining and suggesting further application of FinBERT across different financial NLP tasks. FinBERT's utility could be significantly extended by integrating more extensive stock market datasets, presenting opportunities for more intricate market analysis and model refinement.

Mishev et al. [11] provided evidence that contextual embeddings substantially improve efficiency for sentiment analysis over traditional lexicons and static word encoders, a benefit that holds even in the absence of large datasets. This advancement points to the potential of LLMs to revolutionize sentiment analysis with a more profound understanding of contextual nuances in financial texts.

Deng et al. [66] revealed that LLMs can achieve remarkable outcomes in market sentiment analysis. The study showed that with minimal examples, it is possible to calibrate a 'student' model that matches or surpasses the performance of more extensive, state-of-the-art models, optimizing both effectiveness and computational efficiency.

Fazlija and Harder [64] identified that sentiment scores derived from news content play a critical role in predicting the direction of stock prices. The correlation between news sentiment and market performance underscores the value of high-quality, content-based sentiment indicators in forecasting models.

VII. CASE STUDY REGARDING THE CORRELATION BETWEEN NEWS SENTIMENT AND BITCOIN PRICE

This case study aims to explore the relationship between the sentiment expressed in cryptocurrency news articles and the price fluctuations of Bitcoin. Leveraging the power of sentiment analysis through advanced language models, this study seeks to provide a deeper understanding of how public sentiment, as reflected in media [16], can impact financial markets, particularly the volatile cryptocurrency sector.

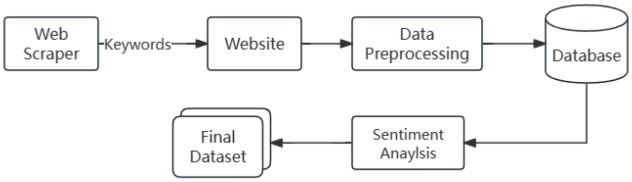

FIGURE 4. Dataset creation process.

A. DATA COLLECTION AND ANALYSIS METHOD

1) CRYPTOCURRENCIES DATA

We collected comprehensive daily cryptocurrencies data from the investing website, www.investing.com,[5] to investigate this relationship. The dataset spans two years, from November 1, 2021, to November 1, 2023, and encompasses various metrics, including price, closing price, highest and lowest price of the day, opening price, and volume of transactions. Consistent with methodologies employed in similar studies [70], the price was chosen as the primary target variable. This decision is based on the Bashchenko [26] price's everyday use as a critical indicator of market sentiment in financial research, providing a reliable measure of the market's end-of-day valuation for Bitcoin.

Given the current existence of approximately 1,000 cryptocurrency coins, some of which suffer from incomplete information or delayed publication, our selection criteria focused on coins with at least 1,000 recorded observations. This threshold ensures the accumulation of sufficient data for our analyses. Importantly, the use of transfer entropy as a methodological approach in our study offers the advantage of not necessitating a balanced dataset, thus allowing for a broader inclusion of data points. Our dataset represents over 80% of the cryptocurrency market's total market capitalization, ensuring a comprehensive analysis scope.

2) NEWS DATA

To gather cryptocurrency-related news data, we employed an open-source Python library, scrape,[6] renowned for its efficiency in web scraping. This tool was instrumental in compiling a substantial dataset of tweets about various cryptocurrencies. To ensure a targeted and relevant data collection, we used a set of carefully selected search keywords for each cryptocurrency. For instance, in the case of Bitcoin, the search parameters included a combination of its name and symbol, such as 'BTC OR BTC OR BITCOIN OR Bitcoin'. Aligning with the timeframe of our price data, the collection period for the news data was also set from November 1, 2021, to November 1, 2023.

Given the sheer volume of cryptocurrency-related news and the constraints of our computational resources, it was necessary to limit the quantity of news collected daily for each cryptocurrency. Without such a limit, the sentiment analysis process would have been impractically prolonged,

[5] https://www.investing.com/crypto/bitcoin/btc-usd
[6] https://github.com/JustAnotherArchivist/snscrape

TABLE 4. **Summary of literature review.**

Reference	Objective	Results Highlights	Research Gap
Zou and Herremans [61], 2023	The impact of embedding actual Tweets informed by domain knowledge for predicting Bitcoin price.	Enhanced accuracy in cryptocurrency price forecasts by integrating emotions and sentiments extracted from Twitter posts into prediction models.	Further model refinement with cryptocurrency-specific terms and evaluation of complex trading strategies are key areas for future exploration.
Raheman et al. [62], 2022	Explore various methods for calculating sentiment metrics from text, assessing their accuracy for cryptocurrency prediction task.	Identification of a model that surpasses over 20 others in performance, offering efficient fine-tuning due to its interpretable nature.	Investigating the predictive power of sentiment connections to improve decentralized finance applications is a prospective research avenue.
Nguyen et al. [27], 2023	Introduce a framework using ChatGPT to assess cryptocurrency market sentiment through social media data analysis.	The empirical suggests that ChatGPT3.5 captures unique aspects of Bitcoin-related sentiment not detected by established indicators.	Future research should apply Generative AI to analyze both company-level data and social media for deeper market insights.
Kulakowski and Frasincar [42], 2023	1. Pre-train and fine-tune a Twitter-oriented model, CryptoBERT, based on BERT architecture. 2. Develop the Language-Universal Cryptocurrency Emoji (LUKE) sentiment lexicon and prediction pipeline, which leverages prevalent emoji sentiment on social media.	The LUKE sentiment lexicon and prediction pipeline have broader applications, such as projecting distant labels for non-English corpora, enabling their direct use in sentiment classification or training other models.	Expanding the lexicon's use to diverse linguistic tasks beyond sentiment classification presents a significant opportunity.
Ortu et al. [63], 2022	Conduct a comprehensive analysis of cryptocurrency price movement predictability using four different deep learning algorithms.	The study finds significant improvement in prediction accuracy across all algorithms when incorporating both trading and social media indicators.	Investigating the magnitude prediction of price movements in addition to directionality is a critical future research area.
Bashchenko [26], 2022	Utilize an advanced NLP algorithm (SBERT network) to embed linguistic data into vector space, enabling intuitive classification.	Demonstrates that over 16% of Bitcoin return variation is explained by fundamental news, challenging the notion that Bitcoin's price is solely driven by speculation.	Incorporating a broader range of news platforms would enhance the study's generalizability.
Kim et al. [60], 2023	Develop CBITS: a Cryptocurrency BERT Incorporated Trading System, which employs pre-trained language models for analyzing Korean cryptocurrency sentiment to improve Bitcoin (BTC) trading strategies.	Successfully adapted a language model to the crypto domain, demonstrating that calculated crypto news sentiments can improve BTC trading models' performance.	Developing strategies for sentiment-based systems to perform in varied market conditions is essential.
Yu et al. [43], 2023	Develop FINMEM, a novel LLM-based agent framework, to enhance financial decision-making by processing multi-source information, establishing reasoning chains, and prioritizing critical tasks.	FINMEM excels in converting diverse financial data into informed investment strategies, highlighted by its proficiency in integrating various data types and reduced training duration, beneficial for trading with new companies.	Exploring FINMEM's application in a multi-agent trading system for portfolio optimization is a promising future direction.
Li et al. [44],2023	Introduce an innovative LLM multi-agent framework with layered memories to more accurately mimic the hierarchical nature of human memory, addressing the challenge of LLMs in prioritizing tasks.	The framework is particularly effective for stock and fund trading, facilitating the extraction of relevant insights from hierarchical financial data to inform trading decisions.	Extending the LLM-based multi-agent design to other sectors for efficiency and collaboration is a promising research direction.
Araci [12], 2019	Propose the use of pre-trained language models, specifically introducing FinBERT (based on BERT), for NLP tasks in the financial domain due to their efficiency with fewer labeled examples and adaptability to domain-specific corpora.	FinBERT proves effective in financial text mining, surpassing other models. It is suggested to broaden its application to additional financial NLP tasks and to include stock market data for more extensive analysis.	Integrating FinBERT with stock market return data and exploring its use in finance-specific tasks like entity recognition is key.
Mishev et al. [11], 2020	Address the challenges of financial sentiment analysis, particularly the domain-specific language and scarcity of large labeled datasets, by designing an evaluation platform to assess the effectiveness of various sentiment analysis approaches.	Developed a platform for evaluating various sentiment analysis methods in finance, combining text representation techniques with machine-learning classifiers.	The research identifies a need for more effective sentiment analysis models specifically tailored to the finance sector, given the inadequacy of general models in this domain.
Fazlija and Harder [64], 2022	Extract financial market sentiment information from news articles to predict the price direction of the Standard & Poor's 500 stock market index.	Demonstrates the usefulness of sentiment scores derived from news content in predicting stock price direction.	Future research could improve prediction quality by expanding data scope and incorporating advanced machine learning models.
Gupta [65], 2023	Simplify the assessment of Annual Reports of various firms by leveraging the capabilities of LLMs.	The walkforward test results indicate promising outperformance relative to S&P500 returns. The research highlights that depending on the target variable, the ML model can outperform S&P 500 benchmark returns.	Future studies could benefit from incorporating a more diverse array of financial data for a comprehensive analysis.
Deng et al. [66]. 2022	Address the challenge of limited high-quality labeled data in financial analysis by employing semi-supervised learning with a Large Language Model (LLM).	Demonstrates that prompting the LLM to create Chain-of-Thought summaries and guiding it through several reasoning paths results in more stable and accurate labels. The use of regression loss further enhances distillation quality.	Addressing the risks of market manipulation and social media influence on financial sentiment is crucial.
Wu et al. [10], 2023	Introduce BloombergGPT, a 50 billion parameter language model trained on a diverse array of financial data.	Validated the model's performance on standard LLM benchmarks, open financial benchmarks, and specific internal benchmarks tailored to its intended usage.	Investigating the reduction of harmful language generation in financial LLMs using cleaner training data is an open question.

potentially taking months. To achieve this, we implemented a timed request mechanism, where news was requested every 20 seconds using a single computer. This approach was crucial to avoid triggering anti-crawler mechanisms on the websites we scraped, while also ensuring a consistent data collection rate. By limiting requests in this manner, we capped the collection at a maximum of 5000 news articles per day for each cryptocurrency, totaling 18506 articles for the entire study period. For each piece of news, we meticulously recorded several key attributes: the date and time of the post (DateTime), the headline, the main text of the news article, the author's information, the URL, and a few other relevant features.

3) SENTIMENT CLASSIFIERS

FinBERT, introduced by Araci in 2019 [12], stands as the first finance domain-specific BERT model, pretrained on the expansive TRC2-financial corpus. This corpus, a specialized subset of Reuters' TRC2, comprises approximately 1.8 million news articles published between 2008 and 2010. Given the scope of this paper, a detailed exploration of the BERT architecture is beyond our purview, but readers are encouraged to consult Araci's original work for a comprehensive understanding.

The FinBERT model underwent further fine-tuning using the Financial Phrase Bank, a resource developed by Malo et al. in 2013 [71], specifically for sentiment classification tasks within the financial domain. FinBERT's performance in financial sentiment analysis tasks showed a notable 15% improvement over generic BERT models [12]. This enhancement in accuracy and the successful application of FinBERT in studies parallel to ours, such as those by Zou and Herremans [61] and Farimani et al. [72], underscored its suitability for our research objectives.

For our study, we opted to employ the FinBERT model in its pre-fine-tuned state, initially configured by Araci [12]. This decision was driven by the nature of our data set, which primarily consists of unlabeled news articles. Further fine-tuning of FinBERT on other labeled datasets was deemed unnecessary, considering it has already been optimized for sentiment classification using the Financial Phrase Bank. By applying this ready-to-use, finely tuned FinBERT model to our news data, we aimed to leverage its advanced capabilities for accurate sentiment analysis in the financial sector without additional training. FinBERT outputs sentiment scores for each news article on a scale from 1 to 10, where 10 indicates a high confidence level in the news positively impacting Bitcoin prices. Finally, the data are stored in a database (Fig. 4).

4) ENGLE'S BIVARIATE DCC-GARCH TECHNIQUE IN FINANCIAL RETURN ANALYSIS

Engle's bivariate Dynamic Conditional Correlation-Extended Generalized Autoregressive Conditional Heteroskedasticity (DCC-GARCH) technique, introduced in 2002, is a cornerstone model for analyzing the co-movement of financial returns. This technique boasts two significant advantages over other variants in the GARCH model family, such as the Baba-Engle-Kraft-Kroner (BEKK) and Constant Conditional Correlation (CCC) models. Firstly, it exhibits a superior capacity to capture time-varying conditional covariance. This is achieved with less computational complexity than the BEKK model, making it more efficient and accessible for complex analyses. Secondly, unlike the CCC model, which assumes constant correlations over time, the DCC model allows for variation, adding flexibility and realism to the analysis.

The bivariate DCC model's simplicity is particularly beneficial in many return series contexts. A key strength is its ability to directly account for heteroscedasticity by calculating Dynamic Conditional Correlations (DCCs) from standardized residuals. As Chiang et al. [73] noted, this approach ensures that the DCCs are free from biases associated with volatility clustering, addressing concerns highlighted by Forbes and Rigobon [74]. Additionally, the DCC model's proficiency in generating accurate, time-varying estimates of volatilities and correlations is invaluable. It capably reflects the latest market news and responds to regime shifts triggered by shocks and crises. This dynamic analysis of correlations over time facilitates more informed asset allocation and hedging decisions.

Implementing the DCC model involves a two-step process to ascertain conditional correlations. A univariate GARCH model is initially estimated for each return series, yielding the conditional variance. Subsequently, dynamic conditional correlations are derived from these standardized residuals. Following the methodology described by Bauwens and Laurent, the model is delineated like this:

$$R_t = \mu_t + \sum_{t}^{0.5} Z_t \tag{1}$$

where the return vector $\boldsymbol{R_t} = (\boldsymbol{r_t^S}, \boldsymbol{r_t^j})'$ and sector indices, $\boldsymbol{r_t^S}$ and the selected alternative investments, $\boldsymbol{r_t^j} * \boldsymbol{\mu_t} = (\boldsymbol{\mu_t^S}, \boldsymbol{\mu_t^j})'$ is the conditional mean process, and $\boldsymbol{Z_t} \overset{iid}{\longrightarrow} N(0, 1)$ is an (2×1) independent identically distributed random variables vector. The conditional covariance matrix $\sum t = D_t C_t D_t$, with the conditional correlation matrix

$$\boldsymbol{C_t} = \left[\rho_t^{S/j}\right] = \text{diag}(\boldsymbol{Q_t})^{-\frac{1}{2}} \boldsymbol{Q_t} \text{diag}(\boldsymbol{Q_t})^{-\frac{1}{2}} \tag{2}$$

and

$$D_t = \text{diag}\left(\sqrt{h_t^S}, \sqrt{h_t^j}\right) \tag{3}$$

where $\sqrt{h_t^S}$ and $\sqrt{h_t^j}$ denote the univariate GARCH variances, The (2×2) symmetric positive matrix $\boldsymbol{Q}_t$ is given by

$$\boldsymbol{Q}_t = (1 - \alpha - \beta)\bar{\boldsymbol{N}} + \alpha \boldsymbol{\eta}_{t-1} \boldsymbol{\eta}'_{t-1} + \beta \boldsymbol{Q}_{t-1} \tag{4}$$

where $\bar{\boldsymbol{C}}$ is the unconditional correlation matrix of standardized innovations $\boldsymbol{\eta}_t$, The added value of the positive scales α

and β is restricted to $\alpha + \beta < 1$. We obtain the DCCs by

$$\boldsymbol{\rho}_t^{S/j} = \frac{\boldsymbol{q}_t^{S/j}}{(\boldsymbol{q}_t^S, \boldsymbol{q}_t^j)^{\frac{1}{2}}} \tag{5}$$

5) TRANSFER ENTROPY

Transfer entropy offers distinct advantages over traditional methods, enhancing its capability to evaluate information flows, as highlighted by Barnett et al. [75]. Unlike conventional econometric models that rely heavily on domain-specific assumptions and constraints, transfer entropy facilitates a non-parametric analysis of time-series data, minimizing the need for extensive presumptions about stochastic processes. Fundamentally, transfer entropy is grounded in econophysics, focusing on quantifying the directional information flow of a variable over time, rooted in information theory. This concept was originally introduced by Shannon in 1948.

$$H_I = -\sum_i p(i) \cdot \log(p(i)) \tag{6}$$

In this context, i denotes a discrete random variable characterized by its probability distribution, $p(i)$, reflecting the various outcomes it may manifest. H is identified as the most effective function for facilitating this transformation, and H_I is known as Shannon entropy. Shannon's [76] seminal work in 1948 established the groundwork for this methodology, focusing on the uncertainty and dynamism in a variable's processes. Subsequently, Kullback and Leibler [77], in 1951, expanded upon this by integrating an additional element, referred to as process J. Notably, the concept of Transfer entropy gains complexity with the inclusion of more variables and values, indicating a broader and more intricate understanding of entropy.

$$h_I(k) = -\sum_i p(i_{t+1}, i_t^{(k)}) \cdot \log(p(i_{t+1}|i_t^{(k)})) \tag{7}$$

To elaborate further, the marginal probability distributions $p(i), p(j)$ and the joint probability distribution $p(i,j)$ are expected to form a stationary time series. This implies that $i_t^{(k)} = (i_t, \ldots, i_{t-k+1})$ represents a sequence of values over time. Similarly, $h_j(l)$ is defined for process J in a comparable manner. Kullback and Leibler [77], in their 1951 work, introduced a broader application of the Markov process to this context.

$$p(i_{t+1}|i_t^{(k)}) = p(i_{t+1}|i_t^{(k)}, j_t^{(k)}) \tag{8}$$

Transfer entropy revolves around the likelihood of one variable obtaining information from its past and from another variable (j_t). This core idea behind 'Transfer entropy' is to quantify the information exchange between two distinct, random variables. Schreiber [78] elucidated this approach, where I and J represent two separate processes. The formula for transfer entropy from J to I is defined as the difference between the information absorbed by a future instance of process $I_{(t+1)}$ from the past values of both I and J, and the information absorbed by the same future instance solely from the past values of I. In essence, transfer entropy seeks to measure the net information flow.

$$T_{J\to I}(k,l) = \sum_{i,j} p(i_{t+1}, i_t^{(k)} j_t^{(l)}) \cdot \log\left(\frac{(i_{t+1}|i_t^{(k)}, j_t^{(k)})}{(i_{t+1}|i_t^{(k)})}\right) \tag{9}$$

In this context, $T_{J\to I}$ is used to assess the flow of information from J to I. Dimpfl and Peter [79] introduced novel methods, including the Markov block bootstrap and the repeated bootstrap, to this field of study. They base their investigation on the null hypothesis which posits the absence of any information transfer.

$$\begin{aligned} &RT_{J\to I}(k,l) \\ &= \frac{1}{1-q} \log\left(\frac{\sum_i \phi_q(i_t^{(k)}) \cdot p^q(i_{t+1}|i_t^{(k)})}{\sum_{i,j} \phi_q(i_t^{(k)}, j_t^{(k)}) \cdot p^q(i_{t+1}|i_t^{(k)}, j_t^{(k)})}\right) \end{aligned} \tag{10}$$

Here, J and I represent two distinct processes, while q is a positive weighting parameter $q > 0$ applied to the individual probability function $p(.)$ for computations. Specifically, i_n refers to the n^{th} element of the time series I, and j_n denotes the $n6th$ element of the time series for the variable J. It should be recognized that $\phi_q(j) = \frac{p^q(j)}{\sum_j p^q(j)}$ and ϕ_q constitute the escort distribution as defined by $\phi_q(i) = \frac{p^q(i)}{\sum_i p_i^q}$. The primary purpose of introducing the Markov process into this analysis is to estimate the likelihood of transitioning from one state to another during information transfer, as well as to facilitate the prediction of potential transition matrix scenarios. According to Equation 10, and following the methodology suggested by Bekiros et al. [80], setting $l = k = 1$ allows for the denoising of the dataset and enables Transfer entropy to detect asymmetrical interactions between pairs (X and Y) and (Y and X), thus offering valuable insights into the dynamics of information flow between two time series. In essence, transfer entropy relies on the logarithmic scale of the number of possible outcomes, determined by a given probability distribution, to analyze information flows.

B. RESULTS

Table 5 presents the descriptive statistics for a case study analyzing the volatility of Bitcoin prices and news sentiment from November 1, 2021, to November 1, 2023. This period witnessed significant fluctuations in Bitcoin prices, as evidenced by a minimum price of \$15,766 on November 21st, 2022, and a peak of \$67,526 on November 8, 2021. Brown [81] suggests that kurtosis values typically range from -10 to $+10$, and skewness values between -3 and $+3$ are acceptable. In this context, the Bitcoin price exhibits a skewness more significant than one and a kurtosis exceeding 3, indicating a distribution with a higher peak and thicker tails than a normal distribution, thereby implying a higher likelihood of extreme values. Regarding news sentiment, the skewness is 0.9311, denoting a moderate positive skew with a longer right tail. The kurtosis of 4.6131, above 3, categorizes the distribution as leptokurtic,

TABLE 5. Descriptive statistics.

Variables	Mean	Std. Dev	Minimum	Maximum	Skewness	Kurtosis
News Sentiment	5.1243	0.9023	2.75	8.8356	0.9311	4.6131
BTC (USD)	29816.38	10877.36	15766	67526	1.1706	3.9919
BNB (USD)	322.6296	100.4686	197.0900	653.7900	1.4066	4.5260
ETH (USD)	2066.9031	873.5813	995.5800	4808.0900	1.4060	4.1125
DOGE (USD)	0.0967	0.0456	0.0528	0.2822	1.7802	5.9480
TRON (USD)	0.0705	0.0121	0.0499	0.1238	1.0853	4.4233
XRP (USD)	0.5423	0.1996	0.3075	1.2818	1.2499	4.1077
SOL (USD)	54.97	56.18	9.66	258.48	1.80	2.35
ADA (USD)	0.59	0.42	0.24	2.27	1.70	2.30

suggesting that the news sentiment scores have a sharper peak and heavier tails compared to a normal distribution.

Table 6 details the outcomes of unit root tests conducted on the variables utilized in this study, explicitly presenting the Augmented Dickey-Fuller (ADF) test results for each factor. Despite the fact that stationarity is not a prerequisite for utilizing the transfer entropy approach, which can handle probability density functions from a single realization as highlighted by Wollstadt et al. [82], we nevertheless proceeded to perform a stationarity test.

For the Bitcoin price, a first-order difference was applied. The ADF test result for Bitcoin price, with a statistic of −2.4729 and a higher p-value, indicates that the time series is non-stationary. This means that the null hypothesis of a unit root for Bitcoin price cannot be rejected at the 5% significance level, necessitating further analysis due to its non-stationary nature.

In contrast, the ADF test for News Sentiment yields a statistic of −6.0595 with a p-value of 0.01. This p-value, being below the commonly accepted significance level of 0.05, strongly refutes the null hypothesis of a unit root. Consequently, we can confidently reject the null hypothesis, affirming that the news sentiment time series is stationary.

Other cryptocurrencies show mixed results: **BNB** has an ADF test statistic of −2.9179, indicating non-stationarity as the null hypothesis cannot be rejected. **ETH** has an ADF test statistic of −2.4665, which is non-stationary. **DOGE** shows a statistic of −3.7383, which rejects the null hypothesis at the 5% significance level, indicating stationarity. **TRON** has an ADF test statistic of −2.7859, indicating non-stationarity. **XRP** has an ADF test statistic of −2.8493, also indicating non-stationarity. **SOL** and **ADA** show statistics of −2.8546 and −4.3383 respectively, indicating stationarity..

These results suggest that while News Sentiment is stationary, most of the cryptocurrency prices exhibit non-stationary behavior, requiring second-order difference for further analysis.

1) BITCOIN VOLATILITY: RESULTS FROM DCC-GARCH MODEL

The outcome of the DCC-GARCH model is presented in Table 7, illustrating the dynamic adjustments in conditional correlation within a multivariate DCC model's framework, complemented by the GARCH model's volatility insights. These statistical results reveal that the price of Bitcoin and news sentiment generally exhibit a similar directional movement; however, this relationship is notably weak.

TABLE 6. Unit root test results.

Variables	ADF Test
News Sentiment	-6.0595***
BTC	-2.4729
BNB	-2.9179
ETH	-2.4665 *
DOGE	-3.7383**
TRON	-2.7859
XRP	-2.8493
SOL	-2.8546**
ADA	-4.3384***

Note: ***, **, and * denote a 1%, 5%, and 10% level of significance, respectively.

The ρ (Rho) value of 0.1145 indicates a low long-term correlation between Bitcoin prices and news sentiment, suggesting that, on average, they do not move together closely. The α (Alpha) value is 0.00107, which is very small, implying that recent news events exert minimal influence on the immediate volatility of Bitcoin's price. This means that new information or shocks from news have a negligible short-term impact on Bitcoin's volatility.

Conversely, the β (Beta) value is 0.9874, which is quite high, indicating that past volatility trends have a substantial and enduring impact on the volatility of Bitcoin's price. This high Beta value suggests that historical news patterns are a significant factor in the longer-term volatility of Bitcoin.

TABLE 7. Estimation results for the DCC-GARCH model.

	Rho	Alpha	Beta
Bitcoin-Coeff	0.1145	0.00107	0.9874

2) NEWS-INDUCED SPILLOVER EFFECTS IN CRYPTOCURRENCY MARKETS

Transfer entropy values are calculated and detailed in Table 8 and Table 9. It's important to clarify that these values should not be confused with directional or signal

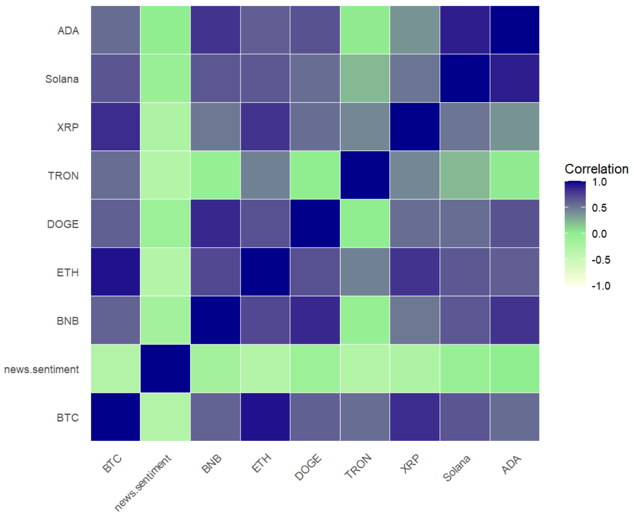

FIGURE 5. **Correlation among cryptocurrencies and news sentiment.**

relationships typical of correlations or coefficients. Instead, they should be understood as transfer entropy measures flowing from the ‘Sender’ to the ‘Receiver’, indicative of the information transfer between the two entities. Our findings highlight significant spillover effects within the cryptocurrency markets, as gauged by the Transfer entropy method.

Notably, cryptocurrencies with smaller market capitalization tend to react more sensitively compared to their larger counterparts. For instance, XRP (ranked 7th) and ADA (ranked 10th) emerge as the most notable recipients. As shown in Table 9, they are the most sensitive to changes, receiving signals from 7 other sources, indicating their high reactivity to market information including news sentiment. BNB, on the other hand, sends signals to 7 other cryptocurrencies, showcasing its role as a significant influencer within the market.

Conversely, cryptocurrencies with the largest market capitalization exhibit lower levels of information exchange. For example, BTC sends information to 7 other cryptocurrencies, illustrating its central role in the market. Despite its significant influence, BTC receives only 4 signals, reflecting its relative stability and lower sensitivity to external shocks compared to smaller cryptocurrencies.

News events that substantially affect Bitcoin’s valuation often initiate a domino effect, impacting the valuations of other cryptocurrencies. For instance, it has been observed that news impacts the prices of BNB, ETH, and XRP. As shown in Table 9, News Sentiment sends 4 shocks to other cryptocurrencies but receives only 1 shock from another cryptocurrency, highlighting its role as a significant source of market information.

Nonetheless, Bitcoin exerts a more profound influence on these cryptocurrencies. Given that Bitcoin accounts for approximately 50% of the total market capitalization of all cryptocurrencies, our observations align with those reported in the study by Zhang et al. [83]. This significant market share underscores Bitcoin’s extensive connectivity and influence over other cryptocurrencies, including BNB, ETH, DOGE, TRON, XRP, SOL and ADA.

C. CASE STUDY LIMITATIONS AND CONSIDERATIONS

The findings of this case study reveal a discernible but modest correlation between news sentiment and Bitcoin price fluctuations. Utilizing the robust FinBERT [12] model for sentiment analysis and the DCC-GARCH technique for financial analysis, we gleaned significant insights into the dynamic interplay between public sentiment, as reflected in media, and Bitcoin’s price volatility. Specifically, the statistical results from the DCC-GARCH model suggested that historical news patterns wield a more substantial impact on Bitcoin’s longer-term volatility than immediate news events. These findings provide insights into the interrelationships between news and Bitcoin price, underscoring the importance of monitoring news for cryptocurrency.

This investigation is subject to certain constraints. Notably, the scope of the data utilized could be more extensive regarding the timeframe and the range of currencies examined, which may influence the perceived relationship between the variables. Including more data and additional cryptocurrencies in future analyses could alter the outcomes of this study. Furthermore, the diversity of methodologies employed in similar studies poses challenges in directly comparing their results. Exploring factors influencing cryptocurrency development is an evolving area of academic interest that warrants further exploration. Future studies aim to overcome these limitations by incorporating broader and more varied datasets and adopting more uniform research methods, contributing to a more cohesive understanding among scholars in the field.

VIII. CHALLENGES

This section delves into the multifaceted challenges and limitations of using LLMs in sentiment analysis. The technical difficulties are paramount, highlighted by the significant computational and storage demands of evolving models like GPT-1 [46] to those with trillions of parameters, raising concerns about accessibility in resource-limited contexts. LLMs also need help with generalizability, often needing help maintaining consistent performance across diverse domains and tasks. This points to a need for models that are more adaptable and versatile. Additionally, the interpretability and ethical usage of LLMs are crucial, especially in critical sectors like finance, where the opaque nature of these models can hinder trust and reliability.

A. CHALLENGES IN LLM APPLICABILITY

The evolution of LLMs has been characterized by a substantial increase in their size, with a progression from GPT-1’s 117 million parameters [46] to GPT-2’s 1.5 billion [45] and a dramatic leap to GPT-3’s 175 billion parameters [47]. More recent models have continued this trend,

TABLE 8. Transfer entropy matrix.

Sender Receiver	News Sentiment	BTC	BNB	ETH	DOGE	TRON	XRP	SOL	ADA
News Sentiment	-	0.0077	0.0061*	0.0074	0.0048	0.0066	0.0087	0.0051	0.0058
BTC	0.0371***	-	0.0019	0.0056	0.0131*	0.0123*	0.0091	0.0290*	0.0069
BNB	0.0102*	0.0455***	-	0.0037***	0.0032	0.0085	0.0034	0.0183***	0.0434***
ETH	0.0215**	0.0427***	0.0181**	-	0.0038	0.0021	0.0058	0.0099	0.0035
DOGE	0.0053	0.0138**	0.0453***	0.0273***	-	0.0064	0.0133*	0.0233***	0.0045
TRON	0.0033	0.0021***	0.0315***	0.0272***	0.0088	-	0.0116*	0.0302	0.0092
XRP	0.0170**	0.0325***	0.0427***	0.0365***	0.0124*	0.0113*	-	0.0194***	0.0025
SOL	0.0054	0.0243***	0.0331***	0.0638***	0.0033	0.0247	0.0097	-	0.0134*
ADA	0.0047**	0.0128*	0.0158***	0.0293***	0.0164*	0.0106*	0.0141*	0.0121*	-

Note: This table presents the transfer entropy between different cryptocurrency pairs. ***, **, and * denote a 1%, 5%, and 10% level of significance, respectively.

TABLE 9. Summary of sending and receiving signals.

Sending		Receiving	
Causal Relationship	Effects	Causal Relationship	Effects
News Sentiment →	4	→ News Sentiment	1
BTC →	7	→ BTC	4
BNB →	7	→ BNB	5
ETH →	6	→ ETH	3
DOGE →	3	→ DOGE	5
TRON →	3	→ TRON	4
XRP →	3	→ XRP	7
SOL →	5	→ SOL	4
ADA →	2	→ ADA	7

reaching into the trillions of parameters [84]. Such vast sizes present formidable challenges regarding storage, memory, and computational requirements. These challenges are particularly acute in scenarios with limited resources or real-time demands, especially when developers cannot access high-powered GPUs or TPUs. For instance, FinBERT is a pre-trained model with 110 million parameters, resulting in a considerable size of 438 MB [12]. The Hugging Face team [85] notes that training a 176 billion parameter model like BLOOM [86] on a 1.5 TB dataset consumes 1,082,880 GPU hours. Similarly, training the GPT-NeoX-20B model [87] on the Pile dataset [88], which includes over 825 GiB of raw text data, requires eight NVIDIA A100-SXM4-40GB GPUs. This extensive training can last up to 1,830 hours or approximately 76 days.

Beyond the monetary costs, these models also incur significant energy expenses. Predictions indicate a massive increase in energy usage by platforms employing LLMs [89], raising environmental concerns. However, a growing body of research is aimed at mitigating these challenges. For example, Wang et al. [90] have demonstrated a distillation method that successfully compresses the MiniLM model to a mere 66 million parameters, significantly reducing its size while maintaining efficiency. Increasing LLM sizes poses a complex challenge, necessitating ongoing efforts for more efficient deployment strategies.

B. CHALLENGES IN LLM GENERALIZABILITY

Generalizability in LLMs pertains to their capability to perform tasks accurately and consistently across various domains, datasets, or functions that differ from their initial training environment. Although LLMs are often trained on extensive datasets, encompassing a broad range of knowledge, their efficacy can be less reliable when applied to unique or niche tasks outside their primary training scope. This limitation becomes evident in diverse applications, from coding projects to document analysis, where the context and semantics can vary significantly across different projects, languages, or domains.

To enhance the generalizability of LLMs, it is crucial to engage in meticulous fine-tuning, apply rigorous validation across diverse datasets, and establish continuous feedback mechanisms. These steps are vital to prevent models from becoming overly specialized in their training data, which can severely restrict their applicability in various real-world scenarios. However, despite these precautions, recent studies indicate that LLMs often need help to extend their high-performance levels to inputs markedly different from their training data [91]. This limitation highlights a significant gap in the current capabilities of LLMs. The challenge, therefore, lies in developing LLMs that possess extensive knowledge and understanding gleaned from large datasets and exhibit the flexibility and adaptability required to function effectively across a wide range of contexts. Addressing this challenge involves refining the training process and innovating in model architecture and learning algorithms.

C. CHALLENGES IN LLM INTERPRETABILITY, TRUSTWORTHINESS, AND ETHICAL USAGE

Interpretability and trustworthiness are pivotal in integrating LLMs for sentiment analysis tasks. The primary challenge lies in demystifying the decision-making processes of these models. Due to their 'black-box' nature, elucidating the mechanisms through which they discern sentiment from text is often challenging. Recent studies [92] have underscored this issue, revealing that while LLMs are proficient in sentiment analysis, their opaque internal workings remain a significant barrier. This obscurity in understanding how these models arrive at their conclusions can generate apprehension and reluctance among users, particularly investors who rely on clear and logical reasoning for decision-making [93].

Investors may only trust the outputs of LLMs with a transparent understanding of the underlying processes.

To foster trust in LLMs, it is essential to develop and implement techniques and tools that shed light on the internal mechanics of these models. Such efforts would enable developers and users to trace and understand the rationale behind the outputs generated by LLMs. Improving interpretability and trustworthiness is a technical necessity and a step towards broader acceptance and use of LLMs in sentiment analysis, leading to more efficient and effective practices in this field [94].

Another aspect contributing to the challenge is the closed nature of many LLMs. Often, it needs to be more transparent about what data these models have been trained on, raising questions about the source training data's quality, representativeness, and ownership. This lack of transparency extends to concerns over the ownership of derivative data produced by the models [95]. Furthermore, the potential vulnerability of LLMs to various adversarial attacks, where inputs are maliciously designed to manipulate or confuse the models, adds another layer of complexity. These risks emphasize the need for robust security measures and ethical considerations in developing and deploying LLMs.

IX. FUTURE OPPORTUNITIES

This section highlights the future opportunities for LLMs in sentiment analysis. As these models evolve and gain prominence in academic research, we explore the emerging trends and potential advancements that could shape their role in sentiment analysis. This section reflects on optimizing LLMs for greater efficiency and effectiveness, expands their natural language processing capabilities to encompass a more comprehensive array of input forms, and discusses enhancing their performance in existing sentiment analysis tasks.

A. OPTIMIZATION OF LLM FOR SENTIMENT ANALYSIS

The ascent of ChatGPT in academic research highlights its growing prominence and acceptance in scholarly circles. Researchers have increasingly favored ChatGPT over other LLMs and their applications since its release, primarily due to its computational efficiency, versatility in handling diverse tasks, and potential for cost-effectiveness [96]. Beyond its application in sentiment analysis, ChatGPT has spearheaded an era of enhanced collaboration in the financial sector. This trend marks a significant shift towards incorporating sophisticated natural language understanding into sentiment analysis [97]. By examining these evolving dynamics, we can anticipate the future trajectory of LLMs like ChatGPT in refining and revolutionizing sentiment analysis processes. These developments indicate the transformative potential of LLMs in sentiment analysis.

Regarding the utilization of LLMs, the choice between using commercially available pre-trained models like GPT-4 and opting for open-source alternatives such as LLaMA [51], LlaMA 2 [52], and Alpaca[7] presents distinct avenues for customization in specialized tasks. The critical difference between these approaches lies in their level of control and personalization. Despite their proprietary nature, pre-trained models like GPT-4 enable quick, task-specific adaptations with minimal data requirements. This approach reduces computational demands and expedites deployment. In contrast, open-source frameworks like LLaMA provide a foundation for extensive tailoring. While these models arrive pre-trained, they can be further adapted, with organizations often modifying and retraining them on large-scale datasets specific to their needs [98]. Although this process demands substantial computational resources and investment, it allows for creating models intricately tailored to specific domains.

B. EXPANDING LLM'S NLP CAPABILITIES IN MORE SENTIMENT ANALYSIS PHASES

Throughout our analysis, it became apparent that most data inputs for LLMs in sentiment analysis were text-based. This finding aligns with traditional NLP approaches, yet there needs to be a noticeable gap in utilizing more diverse and complex datasets, particularly graph-based ones. Embracing a more comprehensive array of natural language inputs, such as spoken language, diagrams, and multimodal data, could significantly expand the capabilities of LLMs in capturing and interpreting varied forms of user sentiment [99].

Integrating spoken language into LLMs could enhance user interactions, enabling the models to process more natural and contextually rich conversations. This addition would allow LLMs to understand better nuances in tone, intonation, and colloquial expressions, which often need to be improved in text-based communication. Similarly, including diagrams could provide valuable visual representations of complex ideas or emotions, offering a unique dimension to sentiment analysis [100]. Diagrams can be a powerful tool to convey information that may be difficult to express through words alone.

Moreover, multimodal inputs that amalgamate text, audio, and visual elements could lead to a more holistic understanding of context. Such a comprehensive approach would likely result in more accurate and context-sensitive sentiment analysis outcomes. For instance, combining textual data with vocal intonations and facial expressions could better understand the user's emotional state and intentions [101], [102].

C. ENHANCING LLM'S PERFORMANCE IN EXISTING SENTIMENT ANALYSIS TASKS

In academic research, establishing a universal and adaptable evaluation framework for LLMs in sentiment analysis is becoming increasingly imperative. Such a framework is essential for conducting systematic and consistent assessments of LLMs, focusing on their performance, efficacy, and potential limitations. This standardization would serve as a critical benchmark, enabling researchers to verify the practical readiness of these models for various applications.

[7] https://github.com/tatsu-lab/stanford_alpaca

A standardized evaluation framework would offer a comprehensive set of criteria and metrics against which to measure LLMs, ensuring that their capabilities are accurately and objectively assessed [103].

In academia, where rigorous analysis and validation are paramount, the absence of such a framework can lead to fragmented and inconsistent evaluations of LLMs, potentially impeding their development and adoption. By establishing a universally accepted framework, researchers can compare different LLMs on a level playing field, fostering a clearer understanding of each model's strengths and areas for improvement. This framework should ideally encompass a range of considerations, including accuracy in sentiment detection, adaptability to different linguistic contexts, computational efficiency, and ethical concerns such as bias and fairness [104], [105].

Furthermore, a universal evaluation framework would facilitate responsible LLM adoption in academic research [106]. It would provide scholars with the tools to decide which models best suit their research needs and objectives.

X. CONCLUSION

In this comprehensive literature review, we adeptly examine the intersection of LLMs and sentiment analysis within financial markets, providing a detailed exploration of LLMs' evolution, application, and future opportunities in this domain. The review navigates through the intricacies of sentiment analysis, underlining its significance in understanding market dynamics and investor behavior. Our meticulous analysis of LLMs, mainly their development from BERT [14] to more sophisticated models like FinBERT [12] and ChatGPT, reveals these models' substantial impact on financial sentiment analysis.

The review methodically dissects the role of LLMs in various financial contexts, from cryptocurrency market prediction to stock price forecasting, showcasing their capability to extract and interpret complex economic sentiments. The case study on Bitcoin price and news sentiment further exemplifies the practical application of LLMs, reinforcing that sentiment analysis, powered by advanced language models, is pivotal in deciphering market trends.

However, the review is open to addressing the challenges and limitations inherent in the current state of LLMs. Issues such as the immense computational requirements, difficulties in generalizability and interpretability, and ethical concerns are thoughtfully discussed, providing a balanced perspective. We call for more efficient deployment strategies, improved generalizability, and enhanced interpretability is particularly compelling, indicating the need for continued innovation in this field.

Looking to the future, integrating more diverse data types and establishing a universal evaluation framework are essential steps toward enhancing the efficacy of LLMs in sentiment analysis. The potential expansion of LLM capabilities to include multimodal data inputs and the implementation of a standard evaluation framework are highlighted as promising avenues for research and development.

REFERENCES

[1] M. Baker and J. Wurgler, "Investor sentiment in the stock market," *J. Econ. Perspect.*, vol. 21, no. 2, pp. 129–152, 2007.

[2] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *J. Finance*, vol. 62, pp. 1139–1168, Jun. 2007. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.2007.01232.x

[3] L. A. Smales, "The importance of fear: Investor sentiment and stock market returns," *Appl. Econ.*, vol. 49, no. 34, pp. 3395–3421, Jul. 2017. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00036846.2016.1259754

[4] T. Rao and S. Srivastava. (2012). *Analyzing Stock Market Movements Using Twitter Sentiment Analysis*. [Online]. Available: http://dx.doi.org/10.1109/ASONAM.2012.30 and https://repository.lincoln.ac.uk/articles/conference_contribution/Analyzing_stock_market_movements_using_Twitter_sentiment_analysis/25165223/2?file=44450105

[5] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Comput. Intell. Mag.*, vol. 9, no. 2, pp. 48–57, May 2014.

[6] V. Ramiah, X. Xu, and I. A. Moosa, "Neoclassical finance, behavioral finance and noise traders: A review and assessment of the literature," *Int. Rev. Financial Anal.*, vol. 41, pp. 89–100, Oct. 2015.

[7] F. Wu, Y. Huang, and Y. Song, "Structured microblog sentiment classification via social context regularization," *Neurocomputing*, vol. 175, pp. 599–609, Jan. 2016.

[8] T. Al-Moslmi, S. Gaber, M. Albared, and N. Omar. (2016). *Feature Selection Methods Effects on Machine Learning Approaches in Malay Sentiment Analysis*. [Online]. Available: https://www.researchgate.net/publication/308968243

[9] R. C. Moore and W. Lewis, "Intelligent selection of language model training data," in *Proc. ACL Conf. Short Papers*, 2010, pp. 220–224.

[10] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "BloombergGPT: A large language model for finance," 2023, *arXiv:2303.17564*.

[11] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: From lexicons to transformers," *IEEE Access*, vol. 8, pp. 131662–131682, 2020.

[12] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," 2019, *arXiv:1908.10063*.

[13] P. Seroyizhko, Z. Zhexenova, M. Z. Shafiq, F. Merizzi, A. Galassi, and F. Ruggeri, "A sentiment and emotion annotated dataset for Bitcoin price forecasting based on Reddit posts," in *Proc. 4th Workshop Financial Technol. Natural Lang. Process. (FinNLP)*, 2022, pp. 203–210. [Online]. Available: https://aclanthology.org/2022.finnlp-1.27

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Oct. 2018, pp. 4171–4186.

[15] J. Kocon, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydlo, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, A. Kocon, B. Koptyra, W. Mieleszczenko-Kowszewicz, P. Milkowski, M. Oleksy, M. Piasecki, L. Radlinski, K. Wojtasik, S. Wozniak, and P. Kazienko, "ChatGPT: Jack of all trades, master of none," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101861.

[16] M. Chakraborty and S. Subramaniam, "Does sentiment impact cryptocurrency?" *J. Behav. Finance*, vol. 24, no. 2, pp. 202–218, Apr. 2023. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/15427560.2021.1950723

[17] A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting information from financial text," *Contemp. Accounting Res.*, vol. 40, no. 2, pp. 806–841, May 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/1911-3846.12832

[18] H. Tong, J. Li, N. Wu, M. Gong, D. Zhang, and Q. Zhang, "Ploutos: Towards interpretable stock movement prediction with financial large language model," 2024, *arXiv:2403.00782*.

[19] A. S. George and A. H. George, "A review of ChatGPT AIs impact on several business sectors," *Partners Universal Int. Innov. J.*, vol. 1, no. 1, pp. 9–23, 2023.

[20] N. A. Sharma, A. B. M. S. Ali, and M. A. Kabir, ''A review of sentiment analysis: Tasks, applications, and deep learning techniques,'' *Int. J. Data Sci. Anal.*, pp. 1–38, Jul. 2024, doi: 10.1007/s41060-024-00594-x.

[21] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, ''A review on large language models: Architectures, applications, taxonomies, open issues and challenges,'' *IEEE Access*, vol. 12, pp. 26839–26874, 2024.

[22] B. Chen, Z. Wu, and R. Zhao, ''From fiction to fact: The growing role of generative AI in business and finance,'' *J. Chin. Econ. Bus. Stud.*, vol. 21, no. 4, pp. 471–496, Oct. 2023.

[23] M. M. Dong, T. C. Stratopoulos, and V. X. Wang, *A Scoping Review of ChatGPT Research in Accounting and Finance*, T. C. Wang and V. Xiaoqi, Eds., Dec. 2023. [Online]. Available: https://ssrn.com/abstract=4680203 and http://dx.doi.org/10.2139/ssrn.4680203

[24] S. A. Farimani, M. V. Jahan, and A. M. Fard, ''From text representation to financial market prediction: A literature review,'' *Information*, vol. 13, no. 10, p. 466, Sep. 2022.

[25] A. Koshiyama, N. Firoozye, and P. Treleaven, ''Algorithms in future capital markets: A survey on AI, ML and associated algorithms in capital markets,'' in *Proc. 1st ACM Int. Conf. AI Finance*, 2020, pp. 1–8.

[26] O. Bashchenko, ''Bitcoin price factors: Natural language processing approach,'' *SSRN Electron. J.*, vol. 13, pp. 22–48, Mar. 2022. [Online]. Available: https://papers.ssrn.com/abstract=4079091

[27] B. N. Thanh, A. T. Nguyen, T. T. Chu, and S. Ha. (2023). *ChatGPT, Twitter Sentiment and Bitcoin Return*. [Online]. Available: https://papers.ssrn.com/abstract=4628097

[28] B. Kitchenham. (2007). *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. [Online]. Available: https://www.researchgate.net/publication/302924724

[29] B. Kitchenham, L. Madeyski, and D. Budgen, ''SEGRESS: Software engineering guidelines for REporting secondary studies,'' *IEEE Trans. Softw. Eng.*, vol. 49, no. 3, pp. 1273–1298, Mar. 2023.

[30] H. Zhao, Z. Liu, Z. Wu, Y. Li, T. Yang, P. Shu, S. Xu, H. Dai, L. Zhao, G. Mai, N. Liu, and T. Liu, ''Revolutionizing finance with LLMs: An overview of applications and insights,'' 2024, *arXiv:2401.11641*.

[31] K. Du, F. Xing, R. Mao, and E. Cambria, ''Financial sentiment analysis: Techniques and applications,'' *ACM Comput. Surv.*, vol. 56, no. 9, pp. 1–42, Oct. 2024.

[32] M. N. Ashtiani and B. Raahemi, ''News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review,'' *Expert Syst. Appl.*, vol. 217, May 2023, Art. no. 119509.

[33] M. Shanahan, ''Talking about large language models,'' 2022, *arXiv:2212.03551*.

[34] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, ''Chain-of-thought prompting elicits reasoning in large language models,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24824–24837.

[35] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, ''Galactica: A large language model for science,'' 2022, *arXiv:2211.09085*.

[36] J. Hoffmann et al., ''Training compute-optimal large language models,'' 2022, *arXiv:2203.15556*.

[37] J. Xu Zhao, Y. Xie, K. Kawaguchi, J. He, and M. Q. Xie, ''Automatic model selection with large language models for reasoning,'' 2023, *arXiv:2305.14333*.

[38] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, ''Unifying large language models and knowledge graphs: A roadmap,'' 2023, *arXiv:2306.08302*.

[39] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, ''ALBERT: A lite BERT for self-supervised learning of language representations,'' in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–17.

[40] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, ''RoBERTa: A robustly optimized BERT pretraining approach,'' 2019, *arXiv:1907.11692*.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[42] M. Kulakowski and F. Frasincar, ''Sentiment classification of cryptocurrency-related social media posts,'' *IEEE Intell. Syst.*, vol. 38, no. 4, pp. 5–9, Jul. 2023.

[43] Y. Yu, H. Li, Z. Chen, Y. Jiang, Y. Li, D. Zhang, R. Liu, J. W. Suchow, and K. Khashanah, ''FinMem: A performance-enhanced LLM trading agent with layered memory and character design,'' 2023, *arXiv:2311.13743*.

[44] Y. Li, Y. Yu, H. Li, Z. Chen, and K. Khashanah, ''TradingGPT: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance,'' 2023, *arXiv:2309.03736*.

[45] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, ''Language models are unsupervised multitask learners,'' *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019. [Online]. Available: https://github.com/codelucas/newspaper

[46] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, *Improving Language Understanding by Generative Pre-training*. Accessed: Feb. 3, 2024. [Online]. Available: https://gluebenchmark.com/leaderboard

[47] T. B. Brown et al., ''Language models are few-shot learners,'' in *Proc. NIPS*, 2020, pp. 1877–1901. [Online]. Available: https://commoncrawl.org/the-data/

[48] J. Achiam et al., ''GPT-4 technical report,'' 2023, *arXiv:2303.08774*.

[49] G. Fatouros, J. Soldatos, K. Kouroumali, G. Makridis, and D. Kyriazis, ''Transforming sentiment analysis in the financial domain with ChatGPT,'' *Mach. Learn. Appl.*, vol. 14, Dec. 2023, Art. no. 100508.

[50] B. Zhang, H. Yang, and X.-Y. Liu, ''Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models,'' 2023, *arXiv:2306.12659*.

[51] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, ''LLaMA: Open and efficient foundation language models,'' 2023, *arXiv:2302.13971*.

[52] H. Touvron et al., ''Llama 2: Open foundation and fine-tuned chat models,'' 2023, *arXiv:2307.09288*.

[53] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, and J. Huang, ''PIXIU: A large language model, instruction data and evaluation benchmark for finance,'' 2023, *arXiv:2306.05443*.

[54] B. Peng, E. Chersoni, Y.-Y. Hsu, L. Qiu, and C.-R. Huang, ''Supervised cross-momentum contrast: Aligning representations with prototypical examples to enhance financial sentiment analysis,'' *Knowl.-Based Syst.*, vol. 295, Jul. 2024, Art. no. 111683.

[55] C. He, C. Li, T. Han, and L. Shen, ''Assessing and enhancing LLMs: A physics and history dataset and one-more-check pipeline method,'' in *Proc. Int. Conf. Neural Inf. Process.*, 2024, pp. 504–517.

[56] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, and W. Y. Wang, ''FinQA: A dataset of numerical reasoning over financial data,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 3697–3711.

[57] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, ''FinBERT: A pre-trained financial language representation model for financial text mining,'' in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 4513–4519. [Online]. Available: http://commoncrawl.org/

[58] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, ''Good debt or bad debt: Detecting semantic orientations in economic texts,'' *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 4, pp. 782–796, Apr. 2014.

[59] K. Cortis, A. Freitas, T. Daudert, M. Huerlimann, M. Zarrouk, S. Handschuh, and B. Davis, ''SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news,'' in *Proc. 11th Int. Workshop Semantic Eval. (SemEval)*, 2017, pp. 519–535.

[60] G. Kim, M. Kim, B. Kim, and H. Lim, ''CBITS: Crypto BERT incorporated trading system,'' *IEEE Access*, vol. 11, pp. 6912–6921, 2023.

[61] Y. Zou and D. Herremans, ''PreBit—A multimodal model with Twitter FinBERT embeddings for extreme price movement prediction of Bitcoin,'' *Expert Syst. Appl.*, vol. 233, Dec. 2023, Art. no. 120838.

[62] A. Raheman, A. Kolonin, I. Fridkins, I. Ansari, and M. Vishwas, ''Social media sentiment analysis for cryptocurrency market prediction,'' 2022, *arXiv:2204.10185*.

[63] M. Ortu, N. Uras, C. Conversano, S. Bartolucci, and G. Destefanis, ''On technical trading and social media indicators for cryptocurrency price classification through deep learning,'' *Expert Syst. Appl.*, vol. 198, Jul. 2022, Art. no. 116804.

[64] B. Fazlija and P. Harder, ''Using financial news sentiment for stock price direction prediction,'' *Mathematics*, vol. 10, no. 13, p. 2156, Jun. 2022. [Online]. Available: https://www.mdpi.com/2227-7390/10/13/2156/htm

[65] U. Gupta, ''GPT-InvestAR: Enhancing stock investment strategies through annual report analysis with large language models,'' 2023, *arXiv:2309.03079*.

[66] X. Deng, V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky, "What do LLMs know about financial markets? A case study on Reddit market sentiment analysis," 2022, *arXiv:2212.11311*.

[67] H. Q. Abonizio, E. C. Paraiso, and S. Barbon, "Toward text data augmentation for sentiment analysis," *IEEE Trans. Artif. Intell.*, vol. 3, no. 5, pp. 657–668, Oct. 2022.

[68] D. Ider and S. Lessmann, "Forecasting cryptocurrency returns from sentiment signals: An analysis of BERT classifiers and weak supervision," 2022, *arXiv:2204.05781*.

[69] J. de Curtò, I. de Zarzà, G. Roig, J. C. Cano, P. Manzoni, and C. T. Calafate, "LLM-informed multi-armed bandit strategies for non-stationary environments," *Electronics*, vol. 12, no. 13, p. 2814, Jun. 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/13/2814/htm

[70] M. Fernandes, S. Khanna, L. Monteiro, A. Thomas, and G. Tripathi, "Bitcoin price prediction," in *Proc. Int. Conf. Adv. Comput., Commun., Control (ICAC3)*, Dec. 2021, pp. 1–4.

[71] P. Malo, A. Sinha, P. Takala, P. Korhonen, and J. Wallenius, "Good debt or bad debt: Detecting semantic orientations in economic texts," 2013, *arXiv:1307.5336*.

[72] S. A. Farimani, M. V. Jahan, A. M. Fard, and S. R. K. Tabbakh, "Investigating the informativeness of technical indicators and news sentiment in financial market price prediction," *Knowl.-Based Syst.*, vol. 247, Jul. 2022, Art. no. 108742.

[73] T. C. Chiang, B. N. Jeon, and H. Li, "Dynamic correlation analysis of financial contagion: Evidence from Asian markets," *J. Int. Money Finance*, vol. 26, no. 7, pp. 1206–1228, Nov. 2007.

[74] K. J. Forbes and R. Rigobon, "No contagion, only interdependence: Measuring stock market comovements," *J. Finance*, vol. 57, no. 5, pp. 2223–2261, Oct. 2002. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/0022-1082.00494

[75] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for Gaussian variables," *Phys. Rev. Lett.*, vol. 103, no. 23, Dec. 2009, Art. no. 238701.

[76] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[77] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[78] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, no. 2, pp. 461–464, Jul. 2000.

[79] T. Dimpfl and F. J. Peter, "Using transfer entropy to measure information flows between financial markets," *Stud. Nonlinear Dyn. Econometrics*, vol. 17, no. 1, pp. 85–102, 2013.

[80] S. Bekiros, D. K. Nguyen, L. S. Junior, and G. S. Uddin, "Information diffusion, cluster formation and entropy-based network dynamics in equity and commodity markets," *Eur. J. Oper. Res.*, vol. 256, no. 3, pp. 945–961, Feb. 2017.

[81] T. A. Brown, *Confirmatory Factor Analysis for Applied Research*. NY, USA: Guilford publications, 2015.

[82] P. Wollstadt, M. Martínez-Zarzuela, R. Vicente, F. J. Díaz-Pernas, and M. Wibral, "Efficient transfer entropy analysis of non-stationary neural time series," *PLoS ONE*, vol. 9, no. 7, Jul. 2014, Art. no. e102833.

[83] H. Zhang, H. Hong, Y. Guo, and C. Yang, "Information spillover effects from media coverage to the crude oil, gold, and Bitcoin markets during the COVID-19 pandemic: Evidence from the time and frequency domains," *Int. Rev. Econ. Finance*, vol. 78, pp. 267–285, Mar. 2022.

[84] S. Moss, "Google brain unveils trillion-parameter AI language model, the largest yet," Tech. Rep., 2021.

[85] S. Bekman, "The technology behind Bloom training," Tech. Rep., 2022.

[86] T. L. Scao et al., "BLOOM: A 176B-parameter open-access multilingual language model," 2022, *arXiv:2211.05100*.

[87] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, and S. Weinbach, "GPT-NeoX-20B: An open-source autoregressive language model," 2022, *arXiv:2204.06745*.

[88] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, "The Pile: An 800 GB dataset of diverse text for language modeling," 2021, *arXiv:2101.00027*.

[89] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, and U. Sauerland, "Risks and benefits of large language models for the environment," *Environ. Sci. Technol.*, vol. 57, no. 9, pp. 3464–3466, Mar. 2023. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.est.3c01106

[90] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2020, 2020, pp. 5776–5788.

[91] A. Albalak, A. Shrivastava, C. Sankar, A. Sagar, and M. Ross, "Data-efficiency with a single GPU: An exploration of transfer methods for small language models," 2022, *arXiv:2210.03871*.

[92] X. Deng, V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky, "What do LLMs Know about financial markets? A case study on Reddit market sentiment analysis," in *Proc. ACM Web Conf.*, 2022, pp. 107–110. [Online]. Available: https://dl.acm.org/doi/10.1145/3543873.3587324

[93] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing reasoning and acting in language models," 2022, *arXiv:2210.03629*.

[94] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, and D. Dou, "Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond," *Knowl. Inf. Syst.*, vol. 64, no. 12, pp. 3197–3234, Dec. 2022. [Online]. Available: https://link.springer.com/article/10.1007/s10115-022-01756-8

[95] S. Sinha, H. Chen, A. Sekhon, Y. Ji, and Y. Qi, "Perturbing inputs for fragile interpretations in deep natural language processing," in *Proc. 4th BlackboxNLP Workshop Analyzing Interpreting Neural Netw. NLP*, 2021, pp. 420–434. [Online]. Available: https://aclanthology.org/2021.blackboxnlp-1.33

[96] M. T. R. Laskar, M. S. Bari, M. Rahman, M. A. H. Bhuiyan, S. Joty, and J. X. Huang, "A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 431–469.

[97] M. U. Haque, I. Dharmadasa, Z. T. Sworna, R. N. Rajapakse, and H. Ahmad, "'I think this is the most disruptive technology': Exploring sentiments of ChatGPT early adopters using Twitter data," 2022, *arXiv:2212.05856*.

[98] I. Gur, O. Nachum, Y. Miao, M. Safdari, A. Huang, A. Chowdhery, S. Narang, N. Fiedel, and A. Faust, "Understanding HTML with large language models," 2022, *arXiv:2210.03945*.

[99] X. Wang, J. He, Z. Jin, M. Yang, Y. Wang, and H. Qu, "M2Lens: Visualizing and explaining multimodal models for sentiment analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 802–812, Jan. 2022.

[100] H. Song, J. Li, Z. Xia, Z. Yang, and X. Du, "Multimodal sentiment analysis based on pre-LN transformer interaction," in *Proc. IEEE 6th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, vol. 6, Mar. 2022, pp. 1609–1613.

[101] K. Dashtipour, M. Gogate, E. Cambria, and A. Hussain, "A novel context-aware multimodal framework for Persian sentiment analysis," *Neurocomputing*, vol. 457, pp. 377–388, Oct. 2021.

[102] U. Sehar, S. Kanwal, K. Dashtipur, U. Mir, U. Abbasi, and F. Khan, "Urdu sentiment analysis via multimodal data mining based on deep learning algorithms," *IEEE Access*, vol. 9, pp. 153072–153082, 2021.

[103] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "ASA: A framework for Arabic sentiment analysis," *J. Inf. Sci.*, vol. 46, no. 4, pp. 544–559, Aug. 2020. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0165551519849516

[104] Z. Ke, J. Sheng, Z. Li, W. Silamu, and Q. Guo, "Knowledge-guided sentiment analysis via learning from natural language explanations," *IEEE Access*, vol. 9, pp. 3570–3578, 2021.

[105] Q. Zhang, J. Zhou, Q. Chen, Q. Bai, J. Xiao, and L. He, "A knowledge-enhanced adversarial model for cross-lingual structured sentiment analysis," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2022, pp. 1–8.

[106] G. F. N. Mvondo, B. Niu, and S. Eivazinezhad, "Generative conversational AI and academic integrity: A mixed method investigation to understand the ethical use of LLM chatbots in higher education," *SSRN Electron. J.*, 2023. [Online]. Available: https://ssrn.com/abstract=4548263 and http://dx.doi.org/10.2139/ssrn.4548263

CHENGHAO LIU received the B.Sc. degree in software engineering from Jiangxi University of Finance and Economics, in 2020. He is currently pursuing the master's degree with The University of Auckland. His research interests include machine learning and large language model to solve financial problems.

ARUNKUMAR ARULAPPAN (Member, IEEE) received the B.Tech. degree in information technology from Anna University, Chennai, India, the M.Tech. degree in computer science and engineering from Vellore Institute of Technology (VIT), Vellore, India, and the Ph.D. degree from the Faculty of Information and Communication Engineering, Anna University, in 2023. He is an Assistant Professor with the School of Computer Science Engineering and Information Systems (SCORE), VIT University. He is proficient with simulator tools MATLAB, ns-3, Mininet, OpenNet VM, and P4 programming. He is exposed to open source tools, such as OpenStack, Cloudify, OPNFV, and Cloud-Native Computing Foundation (CNCF). His research interests include the cloud-native deployment, SDN, NFV, 5G/6G networks, AI/ML based networking, the Internet of Vehicles, and UAV communications.

RANESH NAHA (Member, IEEE) received the M.Sc. degree in parallel and distributed computing from Universiti Putra Malaysia, and the Ph.D. degree in information technology from the University of Tasmania, Australia. He is a Senior Lecturer of information systems with Queensland University of Technology (QUT). He has authored more than 50 peer-reviewed scientific research articles. His research interests include distributed computing (fog/edge/cloud), the Internet of Things (IoT), AI and ML, software-defined networking (SDN), cybersecurity, and blockchain.

ANIKET MAHANTI (Senior Member, IEEE) received the B.Sc. degree (Hons.) in computer science from the University of New Brunswick, Canada, and the M.Sc. and Ph.D. degrees in computer science from the University of Calgary, Canada. He is a Senior Lecturer (an Associate Professor) of computer science with The University of Auckland, New Zealand. His research interests include network science, distributed systems, and internet measurements.

JOARDER KAMRUZZAMAN (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical and electronic engineering from Bangladesh University of Engineering and Technology, Dhaka, and the Ph.D. degree in information systems engineering from the Muroran Institute of Technology, Hokkaido, Japan.

He is a Professor of information technology and the Director of the Centre for Smart Analytics, Federation University Australia. Previously, he was with Monash University, Australia, as an Associate Professor; and Bangladesh University of Engineering and Technology, as a Professor. He has been listed in Stanford's Top 2% Scientists list, since 2020. He has published more than 300 peer-reviewed articles, which include over 110 journals and 180 conference papers. His publications are cited over 7300 times and have an H-index of 36, a g-index of 79, and an i-10 index of 115. He has received over A$5.0m in competitive research funding, including a highly prestigious Australian Research Council Grant and Large Collaborative Research Centre Grants. His research interests include the Internet of Things, machine learning, and cybersecurity. He was a recipient of the Best Paper Award in four international conferences, such as ICICS'15, Singapore; APCC'14, Thailand; IEEE WCNC'10, Sydney, Australia; and IEEE-ICNNSP'03, Nanjing, China. He has served many conferences in leadership capacities, including the program co-chair, the publicity chair, the track chair, and the session chair. Since 2012, he has been an Editor of the *Journal of Network and Computer Applications* (Elsevier). He served as the Lead Guest Editor for *Journal Future Generation Computer Systems* (Elsevier).

IN-HO RA (Member, IEEE) received the Ph.D. degree in computer engineering from Chung-Ang University, Seoul, South Korea, in 1995. From February 2007 to August 2008, he was a Visiting Scholar with the University of South Florida, Tampa, FL, USA. He has been with the School of Computer, Information and Communication Engineering, Kunsan National University, where he is currently a Professor. His research interests include wireless ad hoc and sensor networks, blockchain, the IoT, PS-LTE, and microgrids.