International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I5P125

Volume 72 Issue 5, 204-209, May 2024

Original Article

Real-Time Sentiment Analysis of Twitter Streams for Stock Forecasting

Prabhu Patel

Fellow, Institution of Electronic & Telecommunication Engineers (IETE), New Delhi.

Corresponding Author : prabhu.patel@hotmail.com

Received: 01 April 2024 Revised: 02 May 2024 Accepted: 13 May 2024 Published: 24 May 2024

Abstract *- In an effort to shed light on investor mood and market dynamics, this study explores the use of real-time sentiment analysis of Twitter streams for stock forecasting. The study investigates the potential, methodological developments, practical ramifications, and future directions of sentiment analysis in financial markets through a synthesis of findings from diverse research projects. The summary of results shows sentiment analysis's enormous potential as a forecasting tool, providing insightful information about market patterns and investor mood. Advances in methodology, such as the use of machine learning algorithms and semantic techniques, have greatly improved the precision and efficiency of sentiment analysis models. These developments have opened the door for more resilient and flexible frameworks that can manage streams of data in real-time. Sentiment analysis is a practical tool that traders, investors, product developers, and business managers may use to acquire a competitive edge in the financial markets and inform decision-making processes. However, there are still issues that need to be investigated further, such as data preparation, model optimization, and semantic resource enhancement. The study emphasizes how real-time sentiment analysis in financial markets has the power to alter market dynamics and decision-making processes fundamentally. The predictive capacity of sentiment analysis is poised to transform the financial markets, providing new opportunities and insights for market participants as researchers continue to innovate and improve sentiment analysis approaches.*

Keywords *- Real-Time data streaming, Flink streaming, AI & ML, FinTech.*

1. Introduction

In today's fast-changing financial scene, the ability to accurately predict stock market trends in real time is a highly valued skill (Kolajo et al., 2019). With the rise of social media and the availability of user-generated material, sentiment analysis has emerged as an effective method for evaluating investor mood and projecting market moves. According to Nisar and Yeung (2018), Twitter stream analysis, in particular, has received interest because of its potential to provide immediate insights into market dynamics.

One cannot stress the significance of real-time stock forecasting in the fast-paced trading market of today. Traders and investors look for instruments that can deliver precise and timely insights into market sentiment since markets respond quickly to news, events, and shifts in mood. With Twitter's abundance of user-generated content in real-time, there has never been a better way to gain insightful information about investor sentiment and the collective wisdom of the community (Ibrahim & Wang, 2019).

Furthermore, sentiment analysis has advanced thanks to methodological developments that have made it possible for academics to create more complex and precise predictive models (Yadav & Vishwakarma, 2020; Birjali et al., 2021). These developments include the integration of machine learning algorithms, semantic approaches, and big data analytics frameworks. Conversely, these developments have improved sentiment analysis's predictive power and broadened its useful applications in a number of industries, such as trading, investing, product creation, and company strategy (Birjali et al., 2021).

The results of research that has looked into the predictive ability of sentiment analysis in financial markets are summarized in this review of the literature. It looks into the methods used to analyze sentiment in Twitter data, such as machine learning algorithms, approaches to natural language processing, and semantic approaches. It also looks at the difficulties and restrictions that come with using sentiment analysis, as well as the real-world applications for traders, investors, and financial organizations.

However, there are still a number of obstacles in the way of real-time sentiment analysis for stock predictions. To increase the precision and dependability of predictive models,

a number of important challenges must be overcome, including data quality, noise reduction, sentiment ambiguity, and model resilience. In addition, the dynamic nature of social media platforms and user behaviour demands constant study and adjustment to stay up to date with shifting market conditions.

This review attempts to offer insights into the present level of research in real-time sentiment analysis for stock forecasting by combining data from previous studies. It aims to draw attention to important approaches, conclusions, and difficulties, as well as prospects for additional study and advancement in this emerging discipline. In the end, it seeks to advance knowledge of how sentiment research might be used to improve financial market predictions and decision-making procedures.

2. Methodology

This study used a comprehensive technique to examine existing literature on real-time stock forecasting models. The research methodology included a systematic review of three notable works that made substantial contributions to the discipline.

2.1. Data Collection

Finding and choosing pertinent research on real-time stock forecasting models was a key step in the data collection process. Research on studies that used Twitter data for sentiment analysis was given special attention. Scholarly articles, conference papers, and research publications were found using a methodical search approach across reliable academic sources like PubMed, IEEE Xplore, and Google Scholar.

Search results were filtered using keywords like "real-time stock forecasting," "Twitter sentiment analysis," and "predictive analytics" in order to find papers that matched the goals of the study. To obtain perspectives from practitioners and industry specialists, industry reports, white papers, and blog posts were consulted in addition to academic sources.

2.2. Data Extraction and Analysis

After identifying the pertinent literature, each paper was carefully read and analyzed to obtain important details about the models, procedures, and conclusions. The methods and strategies employed for Twitter data sentiment analysis in real-time and their usage in stock price prediction received particular attention. Important elements, including the performance indicators assessed, the sentiment analysis technique used, and the dataset, were meticulously recorded.

To detect recurring themes, patterns, and trends in the literature, the results from each paper were combined and compared. In order to enable a methodical study of the research findings and insights, this process comprised of classifying and categorizing the data.

2.3. Synthesis of Findings

Synthesizing the results required combining the knowledge gathered from the data extraction and analysis stage with the body of knowledge already available on real-time stock forecasting models. The study aims to provide a thorough overview of the various models used for real-time stock forecasting and their accuracy in predicting stock prices by placing the findings within the larger body of research. To get a comprehensive picture of the research environment, common themes and patterns were found, and variations in techniques and results were compared. Conclusions and implications for potential future study directions in the realm of real-time stock forecasting were derived from the synthesized findings.

3. Literature Review

An essential part of this study is the literature review portion, which offers a thorough examination of the body of knowledge regarding real-time stock forecasting models. This section delves into important research that has made a substantial contribution to our understanding of various strategies and techniques for real-time stock price prediction. Through the synthesis of ideas from a wide range of literature, this review seeks to provide a comprehensive knowledge of the effectiveness, difficulties, and implications of different models used in real-time stock forecasting. Every work that is chosen for review constitutes a distinct contribution to the body of literature, providing insightful information about the approaches, conclusions, and practical consequences.

According to Das et al. (2018) work on "Real-Time Sentiment Analysis of Twitter Streaming Data for Stock Prediction," Rapid technological innovation has led to unparalleled data proliferation in an era that demands efficient administration of massive amounts of real-time data. The need to create reliable data processing architectures that can manage enormous volumes of data in real time has been highlighted by this increase in data output. In light of this, the study's main objective is to forecast stock prices using the Lambda data processing architecture. This architectural framework has gained popularity because it can handle real-time data processing jobs with reduced latency and improved resilience. It is characterized by fault tolerance, extensibility, and scalability.

This study was motivated by the growing need for analysis based on massive datasets, especially in fields like finance, where timely insights can lead to well-informed decisions. Lambda Architecture has gained popularity for real-time stock price prediction due to its effectiveness in handling a variety of use cases, as evidenced by businesses such as Netflix and Yahoo. The ever-expanding information landscape, enabled by social media and financial websites, has transformed conventional market analysis methodologies. As a result, there's a rising interest in using these kinds of data streams to improve the precision and applicability of stock price predictions.

A common foundation for financial data analysis, including stock price prediction, is the efficient market hypothesis, which states that market prices have a random walk pattern because they represent all available information. On the other hand, empirical research indicates that sentiment analysis of social media data can impact stock prices and provide insightful information about market mood. Using massive archives of Twitter data, this study aims to use sentiment analysis to forecast stock values in real time. The study attempts to improve the accuracy of predictive models with each iteration by identifying users' moods and sentiments.

The method used to forecast stock market prices using Lambda Architecture is described in the methodology section. The batch layer, speed layer, and serving layer are the three main layers of this design. While the speed layer processes recent input to minimize latency, the batch layer develops precomputed models and saves immutable past data. For quick and easy retrieval, historical data is indexed by the serving layer. The main data source is the Twitter API, which instantly absorbs streaming data. While Apache Spark uses the JavaScript Object Notation (JSON) format to interpret historical data, Apache Flume makes it easier to input data into the Hadoop Distributed File System (HDFS).

The Stanford Core NLP technology is used to categorize tweets into three categories: neutral, negative, and positive sentiments. Apache Flume ensures reliability and fault tolerance by mediating data transfers between centralized repositories and data generators like Twitter API. The study's conclusions are assessed against actual stock prices that were sourced from Yahoo Finance, and forecasts are contrasted with actual data. The dataset consists of 560,000 tweets over thirteen years, divided into training and testing sets based on the time intervals over which the tweets were gathered. Plotting prediction graphs next to tweets' weighted polarity allows one to evaluate the predictive model's performance.

The findings of the study, Figure 1, show encouraging outcomes when it comes to stock price prediction utilizing sentiment analysis of Twitter data. Yahoo Finance's ground truth stock prices are used as a standard to assess how well the predictive model performs. The dataset offers a thorough foundation for analysis, containing a significant amount of tweets over a period of thirteen years. Prediction graphs show stock price patterns, and extra market sentiment can be gained by examining the weighted polarity of tweets. All things considered, the results highlight the potential of sentiment analysis in real-time stock price prediction, opening the door for additional study and useful financial market applications.

Rajput and Solanki's (2016) work on "Real Time Sentiment Analysis of Tweets Using Machine Learning and Semantic Analysis" emphasized that Social Networking Sites' (SNS) ubiquitous influence has caused an exponential rise in user-generated data, which includes ideas, opinions, and beliefs. Twitter is a prominent social media network with millions of active users who share their opinions on a wide range of topics, including news, events, and products.

Users' thoughts on Twitter not only offer insightful information about how well a product is received and how good it is, but they also act as a gauge of public opinion. Organizations looking to assess public perception and sentiment must employ sentiment analysis as a critical tool since the overall sentiment expressed by people can have a big influence on business decisions.

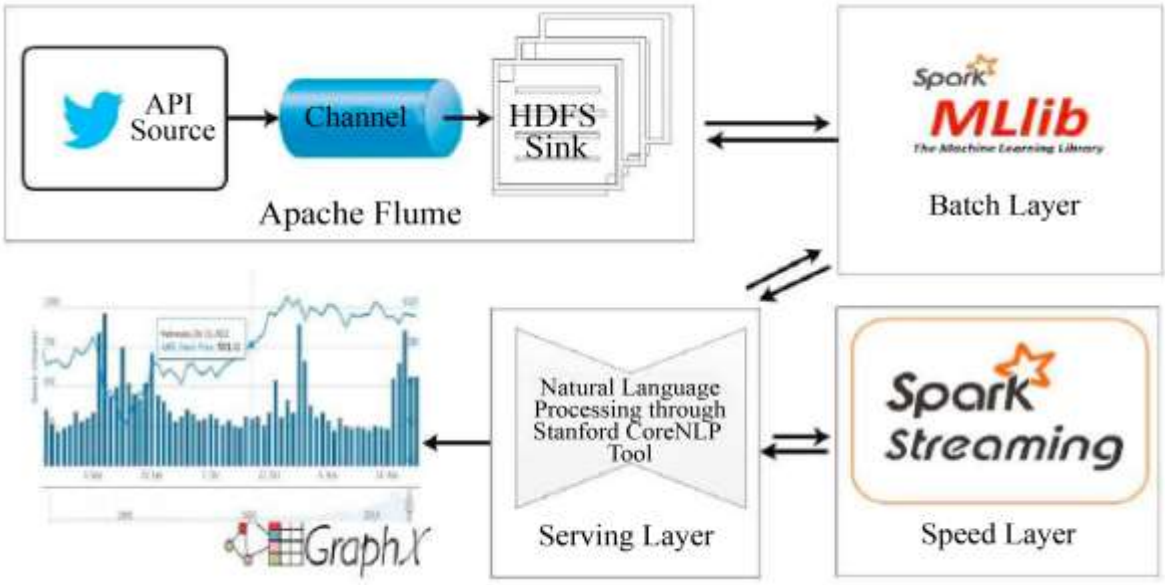

Fig. 1 Lambda architecture for streaming data analysis

Source: Das et al., 2018

Effective public sentiment monitoring has become difficult for organizations due to the vast amount of data that is readily available online. In order to glean valuable insights from the huge amount of user-generated information, automated sentiment analysis algorithms have been developed in response to this difficulty. Sentiment analysis is widely used to forecast stock market trends by utilizing patterns from Twitter. Based on this idea, groundbreaking research by Mishne et al. (2006) showed how sentiment orientations on SNS could be used to predict movie sales, demonstrating the versatility of sentiment analysis in a variety of contexts.

The sentiment analysis of Twitter data for stock market prediction is the goal of both the machine learning technique and the semantic approach, which make up the methodology used in this study. Sentiment analysis is approached from a machine learning perspective, treating it as a text classification issue where classifiers are trained on labelled datasets to predict tweet sentiment. Four classifiers are chosen at the outset of the methodology: Support Vector Machine (SVM), Random Forest, Multinomial Naive Bayes, and Naive Bayes. These classifiers are assessed using measures like Precision, Recall, F-measure, and Accuracy after being trained on a labelled dataset.

Multinomial Naive Bayes stands out as the best option among the classifiers because of its ability to balance speed and accuracy. It performs better than the other classifiers in terms of accuracy and computing efficiency, including Random Forest and SVM. Multinomial Naive Bayes is a perfect fit for sentiment analysis of Twitter data since it works exceptionally well in real-time systems where test data is fetched instantaneously. Bayes classifier can be expressed as in the equation:

$$\log \mathrm{p}(C_k|x) \alpha \log\left(p(C_k)\prod_{i=1}^{n} {p_{ki}}^{x_i}\right)$$

$$= \log p(C_k) + \sum_{i=1}^{n} x_i \,.\log p_{ki}$$

The research uses a semantic technique for sentiment analysis in addition to the machine learning approach. This is using Part-of-Speech (POS) tagging in conjunction with semantic resources like SentiWordNet to classify tweets according to their semantic orientation. Semantic analysis adds to the machine learning approach by offering more context for understanding tweet sentiment.

The study's conclusions demonstrate how well the suggested methodology works to forecast stock market trends using sentiment analysis of Twitter data.

In sentiment analysis of Twitter data, the machine learning method—specifically, the Multinomial Naive Bayes classifier—achieves a high degree of accuracy. Multinomial Naive Bayes beats other classifiers like Random Forest and SVM with an accuracy of 85.06%. Furthermore, it is ideally suited for real-time sentiment analysis when prompt forecasts are essential due to its computational efficiency.

SentiWordNet and POS tagging are incorporated into the semantic approach, which improves sentiment analysis's accuracy. The study obtains an accuracy of 86% by utilizing semantic resources, which enhances the sentiment analysis model's predictive capabilities.

Real-time sentiment analysis of Twitter data is one of the main goals of the study. Real-time data is used to effectively test the suggested algorithm, which achieves up to 77% accuracy. This proves the methodology's viability and efficacy in real-time stock market trend prediction, offering traders and investors insightful information.

In spite of the encouraging outcomes, the study also points out a number of issues that require attention in order to advance, including data preparation and model optimization. Future directions for research include adding more semantic resources and investigating sophisticated machine-learning approaches to improve the sentiment analysis model's robustness and accuracy.

Additionally, Big Data Stream Analytics for Near Real-Time Sentiment Analysis by Otto Cheng and Raymond Lau (2015) claims that user-generated content has proliferated and become omnipresent in the Social Web era. The amount of data produced by people, organizations, governments, and research institutions has increased dramatically; this is known as the "data deluge." With between 100 and 500 million users, online social networking services have emerged as a vital component of personal social contact. Notably, by the end of 2013, the number of active users on Facebook was 1.23 billion, while the number of friendship edges on Twitter was predicted to be over 100 billion. The increase in user-generated material, such as search queries, news articles, online reviews, and private conversations, calls for the creation of sophisticated analytics techniques and systems to handle it efficiently, ideally in real-time or very close to its amount velocity, and variety—the three key components of big data—highlight the necessity for creative solutions to manage the enormous amount and speed of data streams coming from the Social Web. The need to create algorithms that can handle massive data streams in real-time or almost real-time is becoming more and more pressing, even though standard big data analytics algorithms usually work in batch mode.

The idea of a language model, which has its origins in the speech recognition field, has developed to reflect the statistical regularities that underlie language creation. A language model is used in the field of Information Retrieval (IR) to determine

the likelihood that a document may prompt a query. The fundamental unigram language model provides the basis for probabilistic inference in document relevance estimation, which is enhanced by Jelinek-Mercer smoothing. On the other hand, recent research has demonstrated how context-sensitive language models can enhance Information Retrieval (IR) performance. This idea inspires the proposal of an inferential language model, which calculates the likelihood that a term in a document—like a product review—will be located in a Sentiment Lexicon (SL). Environment-sensitive text mining was used to find opinion evidence linked to opinion indicators in an online review environment that is taken into account by this inferential language model. The developed model improves sentiment analysis in dynamic big data streams derived from online social media platforms by enabling context-sensitive opinion scoring.

The findings emphasized that big data analytics research has been booming lately, but there are still not many studies that especially address big data stream analytics. The design and implementation of a novel big data stream analytics framework, BDSASA, specifically suited for the near real-time analysis of consumer attitudes, constitutes one of the research's main theoretical achievements. A probabilistic inferential language model for assessing attitudes inherent in dynamic big data streams from online social media is also demonstrated by the research. The research has practical consequences for product designers and business managers. By utilizing the BDSASA framework, they can enhance their ability to anticipate and analyze consumer preferences for products and services, leading to proactive business strategies. The absence of empirical testing of the suggested paradigm, however, is a drawback of the current work. Future research will concentrate on assessing BDSASA's efficacy and efficiency through the use of social media posts and accurate customer reviews. To improve sentiment polarity prediction, further work will be done to improve the inferential language model and incorporate new data like social network linkages. In order to determine the practical application of the proposed large data stream analytics service, a usability study in actual e-business environments will be carried out.

4. Synthesis of Findings

The combination of research findings from multiple studies shows how sentiment analysis may be used to accurately forecast stock market changes in real-time. Researchers have shown that they can predict market movements and assess investor mood by utilizing the extensive collection of thoughts and feelings shared on Twitter. These promising results suggest that sentiment analysis is a useful tool for understanding market dynamics and investor behaviour, in addition to confirming its feasibility for stock forecasting. For traders and investors looking to obtain a competitive edge, the capacity to use real-time sentiment analysis is becoming more and more crucial as financial markets continue to change in complexity and volatility.

4.1. Methodological Developments

The synthesis highlights important methodological developments in sentiment analysis techniques, especially with regard to the application of machine learning algorithms and semantic techniques. Notable gains in sentiment analysis model accuracy and efficacy have resulted from the use of the Multinomial Naive Bayes classifier and the incorporation of semantic resources like SentiWordNet and POS tagging. These developments not only improve sentiment analysis's accuracy but also help build more resilient and flexible frameworks that can handle the complexities of real-time data streams. In addition, a critical first step in utilizing real-time data analytics for predictive modelling in financial markets is the development of creative large data stream analytics frameworks, such as BDSASA.

4.2. Future Directions and Practical Implications

The synthesized findings have practical consequences for a wide range of stakeholders, including investors, traders, product designers, and business managers. These stakeholders can get actionable knowledge to guide investment strategies, product development projects, and marketing efforts by integrating real-time sentiment analysis insights into decision-making processes. Though the results are promising, they also point to areas that need more research and development. Future research endeavours should prioritize addressing issues related to data preparation, model optimization, and semantic resource augmentation. Furthermore, investigating sophisticated machine learning methods and improving inferential language models offer opportunities to improve the precision and resilience of sentiment analysis systems. Furthermore, real-world empirical testing and usability studies will be necessary to confirm the efficacy and usefulness of these approaches.

The findings summarize the methodological advances, future directions, and transformational potential of real-time sentiment analysis in stock forecasting. Through the use of the abundant information present in social media data streams, academics have the potential to transform financial market decision-making procedures and establish a path for more knowledgeable, flexible, and competitive market involvement.

5. Conclusion

The investigation of Twitter stream sentiment analysis in real-time for stock forecasting opens up new possibilities and innovative possibilities for the financial markets. A summary of the results shows that sentiment analysis is a highly promising forecasting tool that provides insightful information about investor mood and market dynamics. The convergence of big data analytics frameworks, machine learning algorithms, and semantic techniques has enabled substantial methodological advances that improve sentiment analysis models' efficacy and accuracy.

Sentiment analysis holds great potential for stock forecasting as it can use the extensive collection of viewpoints and emotions shared on Twitter to give traders and investors up-to-date information on the mood of the market. Researchers have shown that it is possible to predict market trends and movements by examining the sentiment expressed in tweets.

This gives stakeholders an advantage over competitors in the financial markets and allows them to make well-informed decisions. These results highlight sentiment analysis's revolutionary potential in understanding investor behaviour and broader market dynamics in addition to stock price prediction.

Improvements in sentiment analysis approaches have been essential in raising the efficacy and accuracy of predictive models. The use of machine learning techniques, including the Multinomial Naive Bayes classifier, has allowed researchers to obtain very high sentiment analysis accuracy levels. Furthermore, the use of semantic resources, including SentiWordNet and POS tagging, has improved sentiment analysis models' prediction power through additional refinement. These methodological developments are important steps toward building more resilient and flexible frameworks that can manage the intricacies of real-time data streams.

These discoveries have applications in the fields of trade, investing, product development, and corporate strategy, in addition to theoretical study topics. Stakeholders can gain a competitive edge, predict market trends, and optimize corporate strategy based on consumer preferences and market sentiment by integrating real-time sentiment analysis into decision-making processes. Sentiment analysis provides traders and investors with insightful information about market sentiment, empowering them to take advantage of new trends and make better-informed investing decisions.

Though there is no denying real-time sentiment analysis's potential, there are still issues and room for development. Subsequent investigations ought to concentrate on problems with data pretreatment, model optimization, and semantic resource improvement. Furthermore, verifying the efficacy and usefulness of sentiment analysis approaches will need investigating cutting-edge machine learning strategies and carrying out empirical testing in real-world settings.

The pursuit of real-time sentiment research in financial markets for predicting pulse is, in essence, a dynamic and continuous endeavour. Sentiment analysis's predictive capacity is set to reshape market dynamics and change decision-making processes as researchers continue to push the boundaries of innovation and discovery. This will ultimately shape the future of financial markets.

References

[1] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane, "A Comprehensive Survey on Sentiment Analysis: Approaches, Challenges and Trends," *Knowledge-Based Systems*, vol. 226, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Otto K.M. Cheng, and Raymond Lau, "Big Data Stream Analytics for Near Real-Time Sentiment Analysis," *Journal of Computer and Communications*, vol. 3, no. 5, pp. 189-195, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[3] Sushree Das et al., "Real-Time Sentiment Analysis of Twitter Streaming Data for Stock Prediction," *Procedia Computer Science*, vol. 132, pp. 956-964, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ankur Goel, Jyoti Gautam, and Sitesh Kumar, "Real Time Sentiment Analysis of Tweets Using Naive Bayes," *2016 2nd International Conference on Next Generation Computing Technologies*, Dehradun, India, pp. 257-261, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[5] Noor Farizah Ibrahim, and Xiaojun Wang, "Decoding the Sentiment Dynamics of Online Retailing Customers: Time Series Analysis of Social Media," *Computers in Human Behavior*, vol. 96, pp. 32-45, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Taiwo Kolajo, Olawande Daramola, and Ayodele Adebiyi, "Big Data Stream Analysis: A Systematic Literature Review," *Journal of Big Data*, vol. 6, pp. 1-30, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Tahir M. Nisar, and Man Yeung, "Twitter as a Tool for Forecasting Stock Market Movements: A Short-Window Event Study," *The Journal of Finance and Data Science*, vol. 4, no. 2, pp. 101-119, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] R. Rajput, and A. Solanki, "Real Time Sentiment Analysis of Tweets Using Machine Learning and Semantic Analysis," *The International Conference on Communication and Computing Systems*, pp. 687-692, 2016. [Google Scholar]

[9] Ashima Yadav, and Dinesh Kumar Vishwakarma, "Sentiment Analysis Using Deep Learning Architectures: A Review," *Artificial Intelligence Review*, vol. 53, pp. 4335-4385. [CrossRef] [Google Scholar] [Publisher Link]

[10] Hari Prasad Josyula, "Predictive Financial Insights with Generative AI: Unveiling Future Trends from Historical Data," *Journal of Emerging Technologies and Innovative Research*, vol. 11, no. 1, pp. 354-360, 2024. [CrossRef] [Publisher Link]