AI-Driven Sentiment Analysis of Digital News for Stock Market Prediction

Ms. Sampada Bhonde[1], Mrs. Shilpa Mathur[2], Mr. Abhijeet Jadhav[3], Mr. Kunal Pandey[4]

[1 2 3 4]Assistant Professor

Abstract – Stock Market is a growing field and with developing technology there has been a scope of improvement. With enough insights gained through news feed, articles, tweets of high influence people we can predict the stock market. Artificial Intelligence capabilities can be used to gather insights from textual and even other forms of data such as video and audio data. Therefore, this research paper will delve deeper into the dynamic relationship between the newsfeed and stock market. In an era where news travels rapidly through online platforms and social media, the study employs advanced data analysis techniques to extract meaningful insights from diverse sources, including articles, tweets, and other relevant content. Innovatively, this research introduces a novel approach to news sentiment analysis, incorporating cutting-edge machine learning algorithms and natural language processing tools to enhance predictive capabilities. By exploring previously uncharted territory, this project seeks to identify unique patterns and correlations that have not been systematically examined before. The integration of innovative methodologies not only advances the field of financial forecasting but also provides a fresh perspective on the evolving landscape of information dissemination and its profound implications on global financial markets. The findings of this research contribute to the broader understanding of intersection between technology, information flow, and financial market dynamics.

Keywords: Artificial intelligence, stock market, news sentiment analysis, financial forecasting, correlation analysis, information dissemination, information flow.

1. Introduction

The amount of data and information that is available digitally has skyrocketed since the beginning of the digital era. Not everyone has the time to stay up to date on all the world politics and current affairs, monitoring with which to analyses the optimal stock market move. With the information, even the seasoned brokers may use a little assistance. With the quick development of computer technology, we now possess a powerful tool called artificial intelligence, which is still developing rapidly. The primary conduits for real-time updates and market sentiment are the newsfeeds and social media sites. Considering this, the research discussed here aims to investigate the dynamic relationship between newsfeeds and stock market movements to deliver fresh information that may improve market forecast accuracy.

The research that we're conducting attempts to extract actionable insight from the large and dynamic field of digital information by utilizing the latest data analysis techniques, machine learning algorithms, and natural language processing technologies. This research adopts a comprehensive strategy for analyzing the influence of news on financial markets [9] by focusing on a variety of sources, including articles, tweets, and other online information. This study is unique for its innovative methodology, that presents an innovative perspective on sentiment analysis and looks for distinct patterns and relationships that haven't been thoroughly examined in the body of literature.

The ability to successfully utilize real-time information is becoming increasingly important as financial markets continue to change. With the goal of bridging the gap between traditional financial research and the rapidly changing digital communication landscape, this paper presents a sophisticated understanding of how newsfeeds influence market behavior. Striding into unknown ground, this project aims to contribute to the field of financial forecasting as well as the larger conversation about the interaction of technology, information sharing, and the dynamics of international financial markets. The ultimate goal of this investigation is to identify hidden signals amidst the noise of digital data, providing analysts and investors with useful information to enable better decision-making in an era of abundant information.

2. Literature Survey

Minqing Hu and Bing Liu (2004)

The authors tackled the problem of analyzing large volumes of online customer reviews by proposing methods to extract sentiments, identify product features, and summarize opinions. Their approach efficiently organized user feedback to help both consumers and manufacturers make informed decisions.

Steven Bird, Ewan Klein, and Edward Loper (2008)

Their book introduced the fundamentals of Natural Language Processing (NLP) using Python and the NLTK library. Through real-world examples, they demonstrated key NLP tasks such as text analysis, sentiment detection, and named entity recognition, emphasizing the role of machine learning in language understanding.

Mijana Pejic Bach et al. (2019)

The authors conducted a review on text mining applications in the financial sector, exploring sentiment analysis, topic modeling, and information extraction. They highlighted challenges in managing large financial datasets and stressed the importance of ethical and methodological rigor in financial text analytics.

Mehar Vijh et al. (2019)

This paper reviewed machine learning techniques for predicting stock prices, focusing on text mining methods to extract patterns and insights from financial data. It provided a comprehensive understanding of how AI and big data can improve stock market forecasting and decision-making.

von Beschwitz, Keim, and Massa (2018)

The study examined how quick access to financial news impacts algorithmic trading. It emphasized the advantage of being "first to read" market news and integrating news analytics into trading strategies to enhance performance and decision-making.

Gap Identification: Existing studies lack focus on real-time processing of newsfeeds, multimodal data integration, and personalized investor models. Ethical concerns and the absence of standardized benchmarks also remain underexplored. Addressing these gaps could improve accuracy and fairness in AI-based stock market forecasting.

3. Proposed work

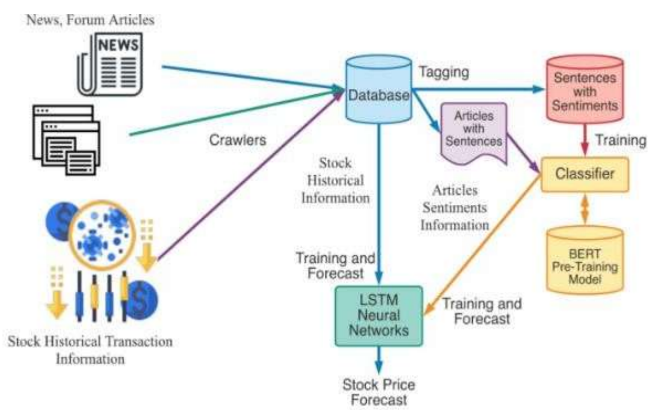

Fig1: Proposed Work

3.1 Web Crawling on News and Forum Articles

Web crawlers collect real-time market data from verified news sites, financial portals, and discussion forums. Using Python tools like Scrapy and BeautifulSoup, structured and multimedia data (text, audio, video) can be extracted, cleaned, and parsed. Real-time pipelines ensure updated information, while databases store the processed data for later analysis.

3.2 Stock Historical Transaction Information

Historical stock data provides a foundation for predictive modeling. After preprocessing (handling missing values, scaling, and encoding), financial indicators such as RSI, MACD, and moving averages are extracted. The data is split for training and testing using models like Linear Regression, Random Forests, SVM, and LSTM, with performance measured by metrics like MSE, Accuracy, and F1-Score.

3.3 Database Operations

Efficient databases manage both historical and live stock data. They support data storage, retrieval, batch updates, and real-time feeds. Predictions are stored alongside actual values for comparison. Logging, access control, and encryption ensure performance tracking and data security.

3.4 Training and Forecasting using LSTM Networks

LSTM networks excel in stock forecasting by learning temporal dependencies and reducing vanishing gradient issues. As shown by Craelius and Daviknes, LSTMs outperform traditional models, effectively capturing both short-term volatility and long-term patterns in financial data.

3.5 Database Tagging to Sentences with Sentiments

Sentiment tagging assigns emotional polarity (positive, negative, neutral) to financial text from news and forums. It enhances prediction accuracy by linking market mood with price fluctuations and supports dynamic tracking of sentiment trends for better investment insights.

3.6 BERT Pre-Trained Model

BERT (Bidirectional Encoder Representations from Transformers) understands financial text contextually by analyzing both directions in a sentence. It improves sentiment detection, entity recognition, and contextual feature extraction. Combining BERT for semantics with LSTM for time-series modeling enhances overall prediction reliability and interpretability.

4. Implementation

4.1 Historical Dataset Training

For training the model on stock historical transactions, two primary datasets are used:

DJIA (Dow Jones Industrial Average) Data: Collected from Yahoo! Finance, this dataset includes the daily open, close, high, and low values. Data is retrieved from historical records to analyze DJIA performance over time, forming a foundational basis for understanding market trends and supporting informed financial decision-making.

Twitter Data: Publicly available Twitter data consists of over 499 million tweets from more than 19 million users [14]. Each tweet includes details such as timestamp, username, and tweet text [3]. Timestamps enable daily segmentation of tweets, aligning with daily prediction and analysis tasks. To ensure data reliability, preprocessing was applied to both datasets. Since DJIA data is unavailable on weekends and holidays when the market is closed, missing values were estimated using a concave interpolation approach. This method predicts missing data points between two available days, ensuring continuity in the dataset. Sharp market fluctuations were smoothed by downward adjustments after steep losses and upward adjustments after spikes to maintain trend stability. Finally, intervals with high inconsistency were removed to prepare the refined dataset for training and testing. This preprocessing ensures the dataset reflects realistic market movements while minimizing noise.

4.2 Sentiment Analysis

Sentiment analysis is a key component of the prediction model, integrating emotional insights from public data. While most research focuses on binary sentiment classification, this study introduces four distinct mood classes — **Mellow, Cheerful, Aware, and Peaceful** — to capture nuanced emotions influencing market behavior.

1) **Word List Development:** The Profile of Mood States (POMS) questionnaire served as the foundation for creating the sentiment lexicon. POMS evaluates moods such as tension, depression, anger, vigor, fatigue, and confusion, which were expanded to suit tweet language.

2) **Filtering Tweets:** To reduce processing load, only tweets likely to express emotions were selected (those containing "feel," "makes me," "I'm," or "I am").

3) **Daily Scoring:** A word-counting technique calculated daily mood scores based on tweet occurrences, normalized by the total number of tweets per day. This simple approach suits short tweet structures.

4) **Mood Mapping:** Each word's score was mapped to one of six POMS states, which were then correlated to the four mood classes. For example, combining vigor and the inverse of depression yielded the score for *Cheerful*.

Granger Causality Analysis was employed to assess how mood variables predict market movements. Statistical significance was determined using p-values — lower values indicated stronger predictive relationships. Results showed that *Cheerfulness* and *Peacefulness* were significant predictors of DJIA movements, particularly when analyzing data from the previous three to four days.

4.3 Technologies and Tools Used

This section outlines the technologies used in the Stock Market Prediction Project, which integrates Twitter sentiment analysis for stock price forecasting. The system leverages tools across data acquisition, NLP, machine learning, and real-time processing.

4.3.1 Data Collection and Preprocessing: • **Python:** Core programming language for data manipulation using Pandas and NumPy. • **Twitter APIs:** Data retrieved using Tweepy for real-time tweet collection.

4.3.2 Sentiment Analysis: • **NLP Libraries:** NLTK and spaCy used for text tokenization, lemmatization, and POS tagging. • **Machine Learning Frameworks:** Scikit-learn and TensorFlow applied for sentiment classification.

4.3.3 Model Development and Deployment: • **Jupyter Notebooks:** Used for interactive analysis and model training. • **Deep Learning:** TensorFlow and Keras frameworks employed to build RNN and LSTM models.

4.3.4 Data Integration and Storage: • **MongoDB:** A NoSQL database used for storing raw tweets, processed sentiments, and prediction results.

4.3.5 Real-Time Monitoring: • **Apache Kafka:** Implements continuous data streaming pipelines for live sentiment tracking.

4.3.6 User Interface and Reporting: • **Flask/Django Frameworks:** Used to develop web dashboards for visualizing stock trends and sentiment analytics.

4.3.7 Deployment and Cloud Services: • **AWS / Google Cloud Platform:** Provides scalable, reliable hosting and deployment infrastructure.

4.3.8 Continuous Improvement and Version Control: • **Git and GitHub:** Facilitate collaborative development, version tracking, and project management.

5. Result and Discussion

5.1 Overview of Sentiment Distribution

The sentiment analysis of the collected newsfeed data revealed a diverse distribution of sentiments. Figure below illustrates the percentage distribution of positive, negative, and neutral sentiments across the dataset. The majority of news articles exhibited a neutral sentiment (46.7%), while positive and negative sentiments were distributed relatively evenly at 43.3% and 10.0%, respectively.

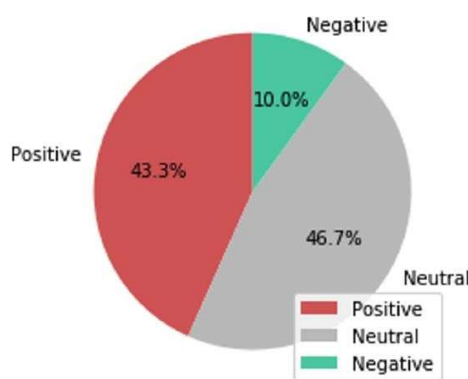

Fig 2: Sentiment Distribution

5.2 LSTM Model for Stock Prediction

Based on the newsfeed data, the Long Short-Term Memory (LSTM) model showed strong performance in stock price prediction. The LSTM model's mean squared error (MSE) was notably lower than that of conventional machine learning models, suggesting that it was successful in identifying subtle patterns and temporal relationships in the dataset.

5.3 Feature Importance and Interpretability

The LSTM model's predictions were significantly influenced by sentiment characteristics extracted from news articles, as demonstrated by feature significance analysis. The BERT pre-trained model's contextual awareness greatly improved the model's capacity to detect minute sentiment nuances, which improved stock price forecasting accuracy.

5.4 Limitations

Although the findings offer valuable insight, it's critical to recognize some limits. The dataset's reliance on news sources could be biased, and stock prices could be impacted by outside variables that were not considered during the analysis. Furthermore, the model's performance could change depending on the state of the market.

Conclusion and Future scope

"Newsfeed The "Newsfeed Analysis for Stock Market Prediction" project utilized advanced machine learning and natural language processing (NLP) techniques to explore how sentiments in news articles influence stock price movements. Through comprehensive sentiment analysis of real-time newsfeed data, a wide spectrum of emotional tones and market sentiments was captured, offering deeper insights into investor behavior and market reactions. The Long Short-Term Memory (LSTM) model outperformed traditional machine learning models for stock price prediction, primarily due to the integration of BERT (Bidirectional Encoder Representations from Transformers) pre-trained models for sentiment understanding. Feature importance analysis revealed that sentiment-based features from financial news significantly enhance stock price forecasting. The contextual depth and linguistic comprehension provided by BERT greatly improved the interpretability and predictive accuracy of the model.

Looking ahead, future research could focus on expanding data sources to include multimedia content such as images, videos, and audio commentary, which may carry additional sentiment cues. Further refinement of sentiment classification methods, evaluation across diverse market conditions, and exploration of event-driven sentiment shifts — such as the impact of major announcements, economic crises, or policy changes — could provide richer, more adaptive prediction models. These enhancements would help bridge the gap between textual emotion dynamics and real-world financial market behavior.

References

[1] Minqing Hu, Bing Liu (August 2004). Mining and Summarizing Customer Reviews.

[2] Steven Bird, Ewan Klein, and Edward Loper. (January 2008). Natural Language Processing in Python

[3] Mijana Pejic Bach, Živko Krstic, Sanja Seljan and Lejla Turulja. (Feb 2019). Text Mining for Big Data Analysis in Financial Sector

[4] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar (Feb 2019). Stock Closing Price Prediction using Machine Learning Techniques.

[5] von Beschwitz, Bastian, Donald B. Keim, and Massimo Massa (2018). First to “Read” the News: News Analytics and Algorithmic Trading

[6] Sanjam Singh, Amandeep Kaur (July 2022). Twitter Sentiment Analysis for Stock Prediction

[7] Faisal Kamiran, Toon Calders (Dec 2011). Data Preprocessing Techniques for classification without discrimination

[8] Enas Al-lozi , Amjed Alfityani, Ayman Abdalmajeed Alsmadi, Amer Moh'd Al_Hazimeh and Jassim Ahmad Al- Gasawneh (June 2022). The role of big data in financial sector: A review paper

[9] Adam Marszk, Ewa Lechman (June 2021). Reshaping financial systems: The role of ICT in the diffusion of financial innovations – Recent evidence from European countries

[10] Robert H. Shumway David S. Stoffer (2006). Time Series Analysis and Its Applications With R Examples

[11] Maureen O’Hara (1995). Market Microstructure Theory

[12] Wen Long, Zichen Lu, Lingxiao Cui (Jan 2019). Deep learning-based feature engineering for stock price movement prediction

[13] Sam Cramer, Michael Kampouridis, Alex A. Freitas (July 2016). Feature engineering for improving financial derivatives-based rainfall prediction.

[14] Johan Bollen, Huina Mao, Xiaojun Zeng (Mar 2011). Twitter mood predicts stock market.

[15] Michael Nofer, Oliver Hinz (June 2015). Using Twitter to predict stock market

[16] Robert Alan Yaffee, Monnie McGee (Jan 2000). Introduction to time series analysis and forecasting: With applications of SASS and SPSS.

[17] Bovas Abraham, Johannes Ledolter (2009). Statistical methods for forecasting