Available online at https://journal.rescollacomm.com/index.php/ijqrm/index

International Journal of Quantitative Research and Modeling

e-ISSN 2721-477X
p-ISSN 2722-5046

Vol. 5, No. 1, pp. 1-11, 2024

Enhancing Stock Trend Prediction Using BERT-Based Sentiment Analysis and Machine Learning Techniques

Nikesh Yadav[1*]

[1]*AVP at Sovrenn Research Lab (SRL), R&D division of Sovrenn, India*

**Corresponding author email: nikesh.yadav@sovrenn.com*

Abstract

Predicting stock trends with precision in the ever-evolving financial markets continues to be a formidable challenge. This research investigates an innovative approach that amalgamates the capabilities of BERT (Bidirectional Encoder Representations from Transformers) for sentiment classification (Pang et al., 2002; ?) with supervised machine learning techniques to elevate the accuracy of stock trend prediction. By harnessing the natural language processing process of BERT and its capacity to understand context and sentiment in textual data, coupled with established machine learning methodologies, we aim to provide a robust solution to the intricacies of stock market prediction. By leveraging BERT's natural language processing capabilities, we extract sentiment features from financial news articles. These sentiment scores, combined with traditional financial indicators, form a comprehensive set of features for our predictive model. We aggregate daily net sentiment, among other metrics, and demonstrate its statistically significant predictive efficacy concerning subsequent movements in the stock market. We employed a machine learning model to establish a quantitative relationship between the aggregation of daily net sentiment and trends in stock market movements. Which improved the state-of-the-art performance by 15 percentage points. This research contributes to the ongoing effort to improve stock trend prediction methods, ultimately aiding market participants in making informed investment choices.

Keywords: Stock Price Prediction, Sentiment Analysis, BERT Model.

1. Introduction

In the ever-evolving landscape of financial markets, the ability to make informed decisions about stock investments is a paramount challenge. Investors and traders continuously seek methods to gain a competitive edge in predicting stock price movements. In this pursuit, advances in natural language processing (NLP) and machine learning have provided new avenues for extracting valuable insights from textual data sources, such as financial news articles, earnings reports, and social media posts.

Sentiment analysis, a branch of NLP, has emerged as a powerful tool for gauging market sentiment and investor emotions by analyzing and classifying the sentiment or emotional tone expressed in textual data. Within this context, the Bidirectional Encoder Representations from Transformers (BERT) model has garnered significant attention. BERT, known for its ability to capture context and semantics in text, has demonstrated remarkable performance in various NLP tasks, including sentiment analysis.

The primary objective of this article is to delve into an innovative fusion of BERT (Bidirectional Encoder Representations from Transformers) with supervised machine learning algorithms in the context of stock price prediction. Our exploration begins by contrasting this novel approach with a more conventional and simplistic method, which entails using a bag-of-words (Joachims, 1998) representation in conjunction with a supervised machine learning algorithm to forecast stock prices. This initial approach serves as our benchmark, providing a foundation against which we can measure the improvements brought about by integrating BERT-based sentiment analysis.

Several methods for stock price prediction have been proposed recently. (Shah et al., 2022) introduced a model for predicting the closing price of the Indian Nifty 50 stock market index. Their approach leveraged Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) to analyze a 20-day window of trading data, extracting features for forecasting the next day's movement.

Aldhyani and Alzahrani (2022) developed an intelligent system that used LSTM and a hybrid CNN-LSTM model to predict closing prices of Tesla and Apple, Inc. based on two years of historical data. Notably, their CNN-LSTM model outperformed both traditional algorithms and standalone LSTM models.

Bansal et al. (2022) addressed stock price prediction for twelve prominent Indian companies. They tested five algorithms, including KNN, Linear Regression (LR), Support Vector Regression (SVR), Decision Tree Regression, and LSTM. By assessing model performance using metrics such as R-Squared Value (R2), Root Mean Square Error (RMSE), and Symmetric Mean Absolute Percentage Error (SMAPE), they found that LSTM consistently delivered superior results.

In a separate study (Lin et al., 2022) focused on forecasting stock indices and closing prices of corporate stocks. Their research underscored the effectiveness of LSTM among various machine learning methods for this task. These recent advancements highlight the evolving landscape of stock price prediction using deep learning and traditional algorithms.

In the initial part of our analysis, we adopt a straightforward yet widely employed technique: the bag of words model. This model involves transforming textual data into a numerical format by counting the frequency of words in a document. We subsequently employ a supervised machine learning algorithm to establish the relationship between these word frequencies and stock price movements. The performance of this initial approach is used as a baseline against which we will compare the enhanced results achieved through the integration of the BERT model.

The subsequent phase of our study introduces the BERT model, a deep learning architecture renowned for its contextual understanding of language. We employ BERT to extract sentiment information from news articles related to the stock market. Unlike the bag-of-words model, BERT captures the intricate contextual nuances of language, providing a more comprehensive and sophisticated representation of the text's emotional tone.

By contrasting these two approaches i.e. traditional bag-of-words with supervised learning and BERT-based sentiment analysis we aim to shed light on the potential enhancements in stock price prediction brought about by the advanced language understanding capabilities of BERT. Our investigation aims to provide valuable insights into the effectiveness of integrating state-of-the-art NLP techniques like BERT into the realm of financial forecasting, ultimately contributing to the evolution of predictive models in the domain of stock market analysis.

In the course of this research endeavor, we endeavor to provide analytical and comprehensive responses to the following research questions

A. To what extent does integrating BERT-based sentiment analysis improve the accuracy of stock trend predictions compared to traditional approaches?
B. How does the choice of supervised machine learning algorithm influence the effectiveness of BERT-enhanced stock trend prediction?
C. What are the limitations and challenges of incorporating BERT sentiment analysis into stock trend prediction models, and how can these challenges be mitigated?
D. How does the temporal aspect of news sentiment affect the predictive power of the BERT-enhanced model, and can real-time sentiment analysis be incorporated for more timely predictions?

BERT (Bidirectional Encoder Representations from Transformers) is a remarkable natural language processing (NLP) model that has redefined the landscape of sentiment analysis. Its working mechanism is a testament to the power of deep learning and pretraining in understanding the nuances of human language.

At its core, BERT is a deep neural network architecture based on the Transformer model, known for its attention mechanism. The key innovation of BERT is its bidirectional nature, which means it considers the context of a word by looking at all the other words in a sentence, both to the left and right. This bidirectionality enables BERT to capture rich contextual information, making it exceptionally effective for understanding the semantics and sentiment expressed in text.

This paper explores the fusion of BERT model sentiment classification with supervised machine learning techniques to address the critical task of stock trend prediction. While traditional financial indicators and technical analysis have been the cornerstone of quantitative analysis in finance, the integration of sentiment analysis through BERT offers a novel dimension to market forecasting. By harnessing the power of deep learning and large-scale language understanding, this approach seeks to uncover the hidden sentiments and market dynamics that can influence stock prices.

The primary objective of this study is to investigate the effectiveness of BERT model sentiment analysis as a feature in stock trend prediction models. Through empirical analysis, we aim to assess whether sentiment-based features, combined with aggregate daily net sentiment, can yield enhanced predictive accuracy. Additionally, we explore the challenges and limitations associated with this integrated approach and highlight potential avenues for future research.

In a world where financial decisions are increasingly influenced by information and sentiment expressed in textual data, this research contributes to the ongoing dialogue on improving stock trend prediction methods. Ultimately, it strives to empower market participants with valuable insights for more informed and data-driven investment strategies.

2. Literature Review

2.1. Bag-of-Words Based Models

Initially, we employ a bag-of-words-based model to identify trends in historical stock data. The financial news, serving as our input data, undergoes an initial phase where it is processed through the bag-of-words mechanism, resulting in a vector representing the word frequency within the text.

In the subsequent step, we apply a masking procedure to the stock data. Specifically, we assign a value of 1 when the closing price is greater than the opening price, and 0 otherwise. To ascertain the connection between our input and output, we employ a variety of machine learning algorithms, including K-Nearest Neighbors (KNN), Logistic Regression (LR), and Support Vector Machines (SVM).

2.2. Fuzzy Bag-of-Words Models

Fuzzy Bag-of-Words (BoW) models represent a notable extension of the traditional Bag-of-Words approach, which is widely employed in natural language processing and information retrieval. The term "fuzzy" in this context alludes to the introduction of a degree of ambiguity and imprecision into the representation of textual data, offering a more nuanced and context-aware approach to language modeling.

In a conventional BoW model, each document is represented as a vector, with word frequencies as its elements. This representation is highly efficient but often lacks the ability to capture the subtle complexities of language, such as word ambiguity, synonyms, or context-based variations. Fuzzy BoW models seek to address these limitations by incorporating a degree of fuzzi- ness or soft clustering into the representation.

One of the key techniques used in fuzzy BoW models is term weighting. Instead of relying solely on strict term frequencies, these models consider term importance based on various factors such as TF-IDF (Term Frequency-Inverse Document Frequency) or semantic relevance. By assigning weights to terms based on their significance within the document and across the corpus, fuzzy BoW models can better capture the nuances of language.

Furthermore, fuzzy BoW models often employ techniques like word embeddings, which represent words as dense vectors in a continuous space. Word embeddings, such as Word2Vec or FastText, encode semantic relationships between words, allowing the model to understand synonyms and word context. This can significantly improve the ability to handle polysemy (multiple words of words) and disambiguate terms in context.

3. Materials and Methods

In the initial phase of our research endeavor, a crucial and systematic step was the acquisition of pertinent news articles, which formed the foundational dataset for our investigation. To ensure the availability of a comprehensive and diverse set of news articles, we implemented a meticulous and well-defined data collection process. The primary objective was to accumulate a substantial repository of news articles spanning a decade, enabling us to explore temporal trends and patterns.

To fulfill this objective, we initiated the process of news article collection by deploying a web scraping methodology. This method facilitates the automated extraction of textual content from a wide array of online sources, encompassing a spectrum of news categories. By employing this technique, we could amass a substantial corpus of news articles, covering various topics, domains, and themes. Each article was stored locally in a secure and organized repository, ensuring data integrity and accessibility throughout the research process.

The temporal scope of our data collection spanned a full decade, offering a unique opportunity to investigate long-term trends and transformations within the news landscape. This comprehensive and meticulously curated dataset of news articles serves as the fundamental bedrock upon which our subsequent analyses, insights, and research findings are constructed, enabling us to gain a profound understanding of the evolving media landscape over this extended time period.

Subsequently, we harnessed the power of Machine Learning (ML) as a pivotal component of our methodology to categorize the acquired news articles into distinct thematic classes. The goal of this categorization process was to discern and classify the extensive array of news articles into specific, well-defined subject categories. This categorization was vital in our quest to analyze the data effectively.

As depicted in the accompanying figure 1, our dataset exhibited a notable diversity in terms of subject matter, encompassing a wide spectrum of topics and domains. To maintain data relevance and ensure the precision of our analysis, we judiciously retained only those articles that are aligned with the primary research objectives, subsequently eliminating non-relevant articles. This meticulous curation of the dataset allowed us to focus our analytical efforts on the most pertinent and informative subset of news articles, thus enhancing the robustness and accuracy of our research outcomes.

Utilizing the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018), we conducted a sentiment analysis on the collected news articles, aiming to discern the prevailing sentiment within the textual content. The

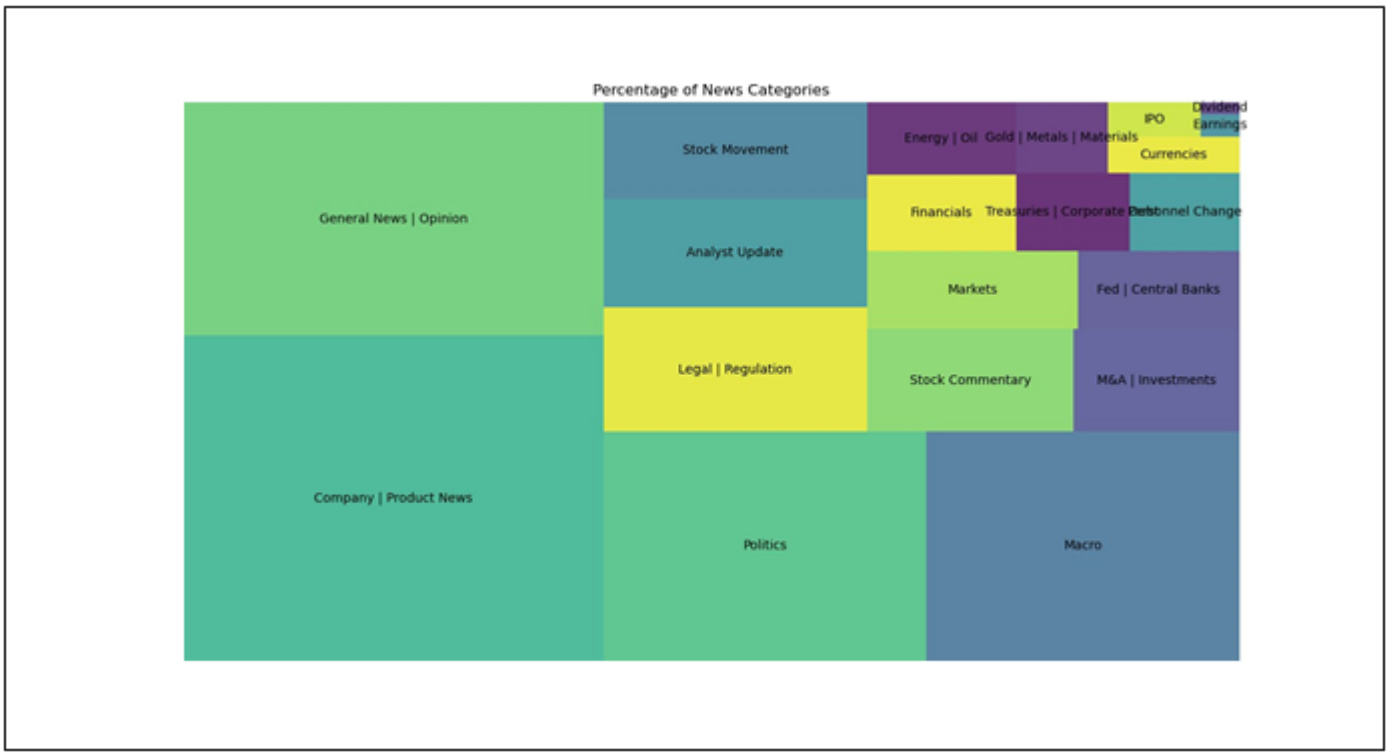

Figure 1 : News articles data-set categories

Date	Article	Category
3/1/2008	Holi in December?	General News | Opinion
3/1/2008	CWG committee for Greener games	Politics
3/1/2008	2008 be the year for Bottom Billion	Macro
1/3/2008	ADB meet South Asia on policy update	Company | Product News
1/3/2008	S&N's 37.5% stake in UB goes to Dutch brewer Heineken	M&A | Investments

BERT model, renowned for its contextual understanding of language, enabled us to classify the sentiment of these articles into three distinct categories: Bearish, Bullish, and Neutral. This classification system provided a nuanced perspective on the articles' sentiment, capturing not only the polarity but also the degree of positivity or negativity. Each sentiment category was associated with a corresponding sentiment score, reflecting the intensity of the sentiment expressed within the article. A representative example of this sentiment classification and scoring can be observed in the tabular data, exemplifying the granularity and depth of sentiment analysis achieved through our methodology.

In the context of our research focused on stock market data collection and analysis, a fundamental and methodical approach was employed to gather pertinent financial information. Specifically, we implemented a web scraping methodology to obtain the historical stock market data pertaining to the NSE50 index. The NSE50, representing the National Stock Exchange's top 50 companies in India, serves as a critical benchmark for gauging the Indian stock market's performance.

Over the course of a decade, spanning the last ten years, we meticulously gathered data encompassing the opening and closing prices of the NSE50 index for each trading day. The meticulousness of our data collection process was paramount, as it ensured that we acquired a comprehensive and high-quality dataset, capturing the long-term dynamics of the market.

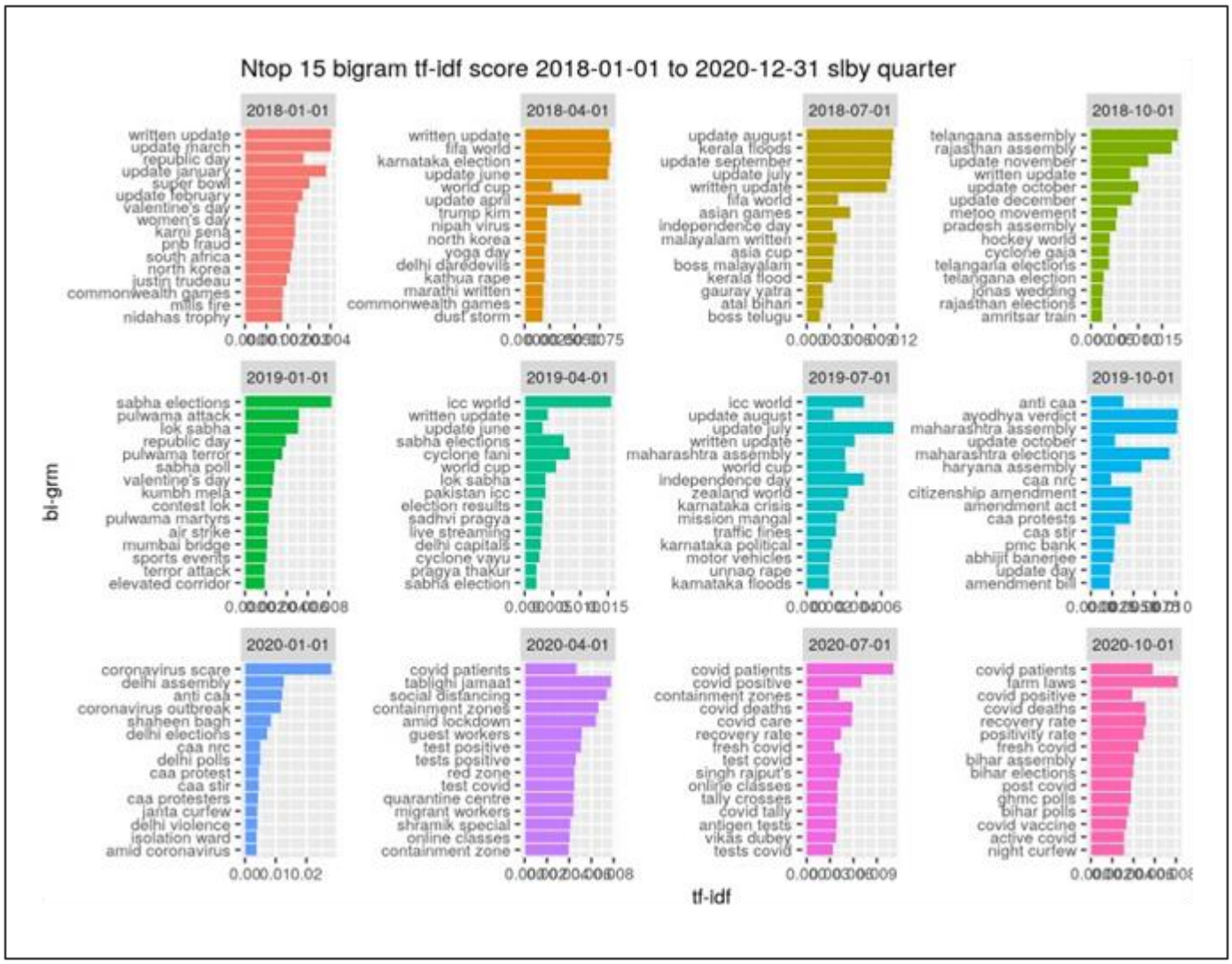

Figure 2 : The figure illustrates the prevalence of news articles topic across various quarters during the period from January 1, 2018, to December 31, 2020

To extract actionable insights from this data, we applied a binary mask. This mask, based on a fundamental market indicator, discriminates between days when the closing price of the index was greater than the opening price (masked as 1) and days when the closing price was lower than the opening price (masked as 0). This binary masking is a common technique used to identify the direction of price movement within the stock market. By doing so, we were able to categorize each trading day into either an "up" or "down" day, facilitating subsequent analyzes to understand market trends, volatility, and potential investment strategies.

The resulting dataset, consisting of these binary masked values, serves as the bedrock of our research, enabling us to explore and uncover significant patterns and trends in the NSE50 index's historical price movements. This methodology provides a robust foundation for the comprehensive examination of stock market dynamics, ultimately contributing to a deeper understanding of financial markets and their inherent complexities.

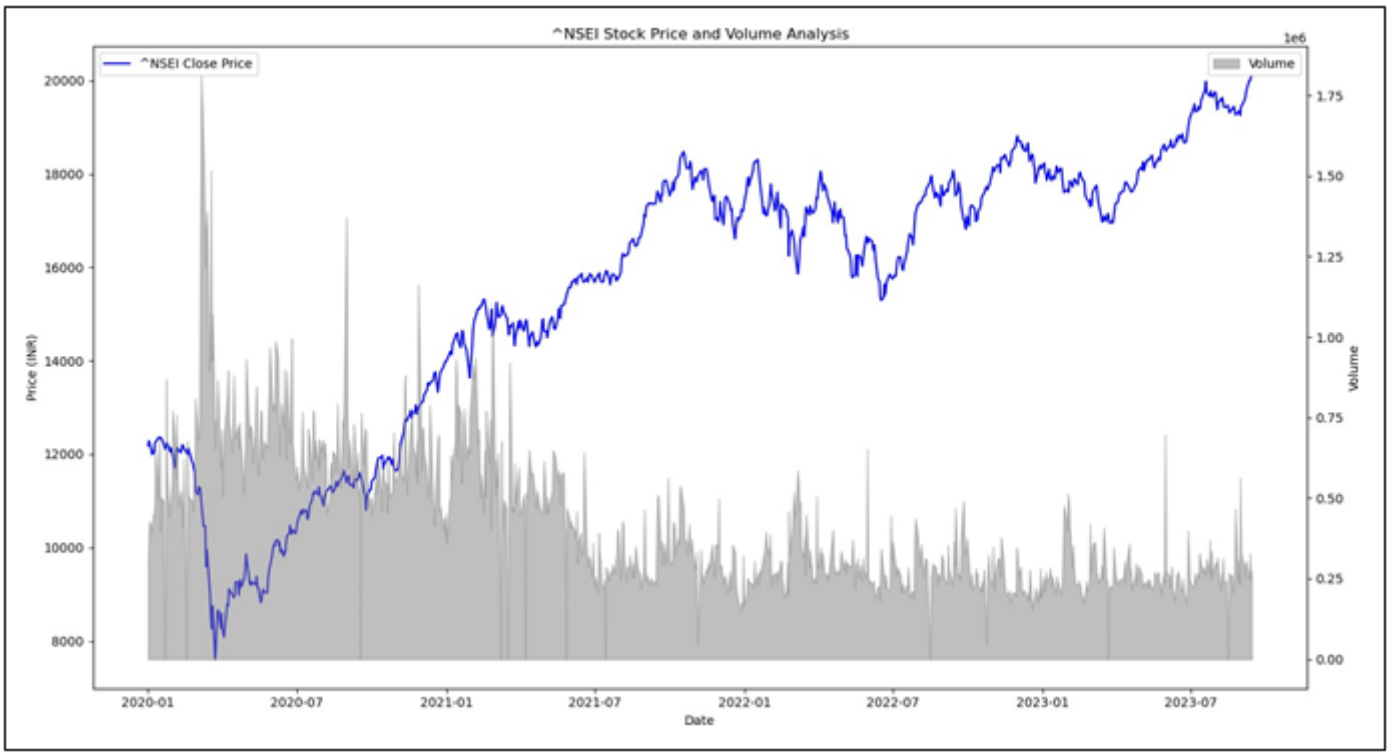

Figure 3 : NIFTY 50 historical data

Article	Sentiment	Score
Gold futures gain on spot demand ,	bullish	0.88
Global tech houses in TN will be given to poor: CM Palaniswami	bearish	0.62
Hyundai reports 33% jump in sales to 66,750 units in December	bullish	0.97
India puts on hold visit request by Ukrainian MPs	neutral	0.64
BJP MLA resigns from party in poll-bound MP	neutral	0.75

4. Results and Discussion

In this section, we present the results and discussion of our study focused on predicting daily stock trends. Our approach involves using the top 20 headlines for each day as our primary data source. To ensure the reliability and generalization of our models, we initially partitioned our dataset into distinct training, testing, and validation sets, adhering to a distribution ratio of 70%, 20%, and 10%, respectively.

We commenced our analysis by employing a combination of Bag of Words (BoW)-based models and various machine learn- ing algorithms, including K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR). This ensemble of models formed the basis for our investigation into the predictive power of sentiment analysis derived from financial news headlines. The subsequent sections provide an in-depth analysis of our findings and the implications they hold for predicting stock trends.

The outcomes of our study, encompassing the performance of each model on our datasets, are meticulously presented in Table 3. This table serves as a comprehensive reference for evaluating the effectiveness of the various machine learning models utilized

Figure 4 : In top portion of Figure variation of NIFTY50 price variation with time is plotted w.r.t time and in the bottom part thecumulative score with respect to time is plotted

in our investigation. Upon careful analysis, it becomes apparent that the performance of all the models consistently falls within a relatively narrow range, with accuracy metrics hovering in the bracket of 50 % to 55%. While this accuracy range is notable and indicative of a certain level of predictive capability, it ultimately falls short of the desired standard for accurately forecasting stock price movements.

The accuracy range of 50% to 55%, although demonstrating a degree of predictive power, remains suboptimal for the complex and dynamic task of stock price prediction. In the financial domain, where even marginal improvements in prediction accuracy can yield substantial gains, such results underscore the inherent challenges of this endeavor. The observed limitations in model performance prompt further scrutiny and emphasize the need for advanced techniques and additional features to enhance predictive precision.

Therefore, our investigation not only highlights the reliability of the models but also underscores the inherent complexities in stock price prediction. It underscores the need to explore more sophisticated strategies, potentially involving the integration of additional data sources, advanced feature engineering, and intricate model architectures, to address the intricacies of financial market dynamics and further elevate predictive accuracy.

In our pursuit of enhancing predictive accuracy, we initiated a multifaceted approach. Our initial step involved the computation of a sentiment score for each financial news article within our dataset. This sentiment score encapsulates the underlying sentiment or tone of the article, whether it leans towards a pessimistic (Bearish) or optimistic (Bullish) outlook.

To further synthesize the sentiment information and streamline our analysis, we amalgamated the Bearish and Bullish sentiment scores, resulting in a Cumulative score. This Cumulative score offered a more comprehensive representation of the overall sentiment associated with the articles

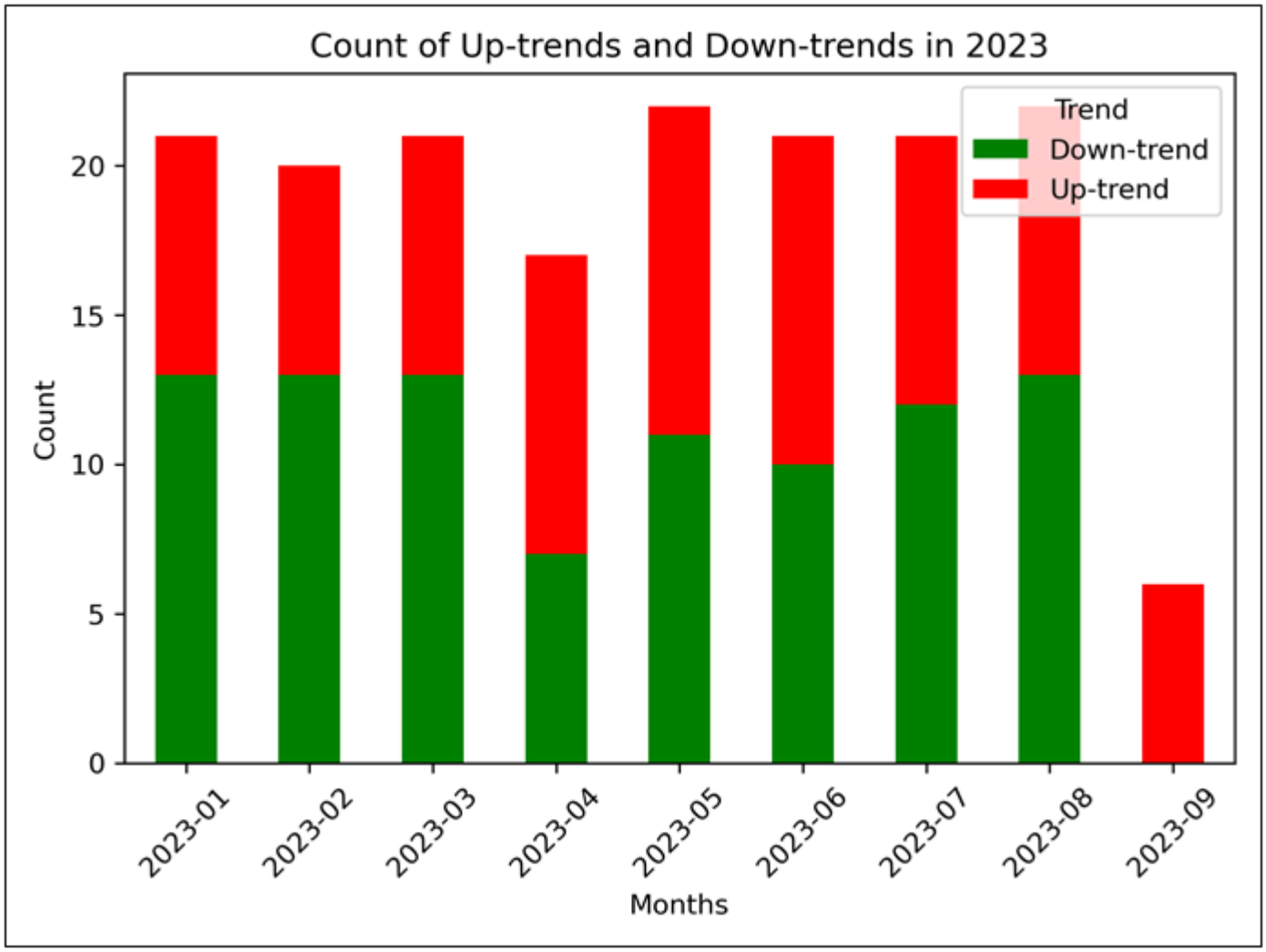

Figure 5 : Figure shows the variation of monthly up-trends and downtrends events in NIFTY 50 price. Blue colors depicts the number of uptrend and red color depicts the number of downtrend in that particular month.

Figure 3 serves as a visual representation of the relationships we explored. In the upper segment of the figure, we depicted the variations in the NIFTY50 stock price over time, providing a crucial context for understanding market dynamics. Meanwhile, in the lower section of the figure, we showcased the evolution of the Cumulative score over the same time frame. This juxtaposition allowed us to visually correlate the shifts in market sentiment, as reflected by the Cumulative score, with the corresponding movements in the NIFTY50 stock price.

The dataset under consideration encompasses a time period characterized by two pivotal events: the outbreak of the COVID-19 pandemic and the Russia-Ukraine conflict. It is noteworthy to observe that during both of these significant events, the bullish sentiment scores consistently reached their nadir, coinciding with pronounced declines in stock prices. This observation underscores the potential for establishing strong correlations between sentiment analysis derived from news articles and market fluctuations through the application of advanced machine learning models. The prospects for extracting deeper insights from this correlation appear promising and warrant further exploration.

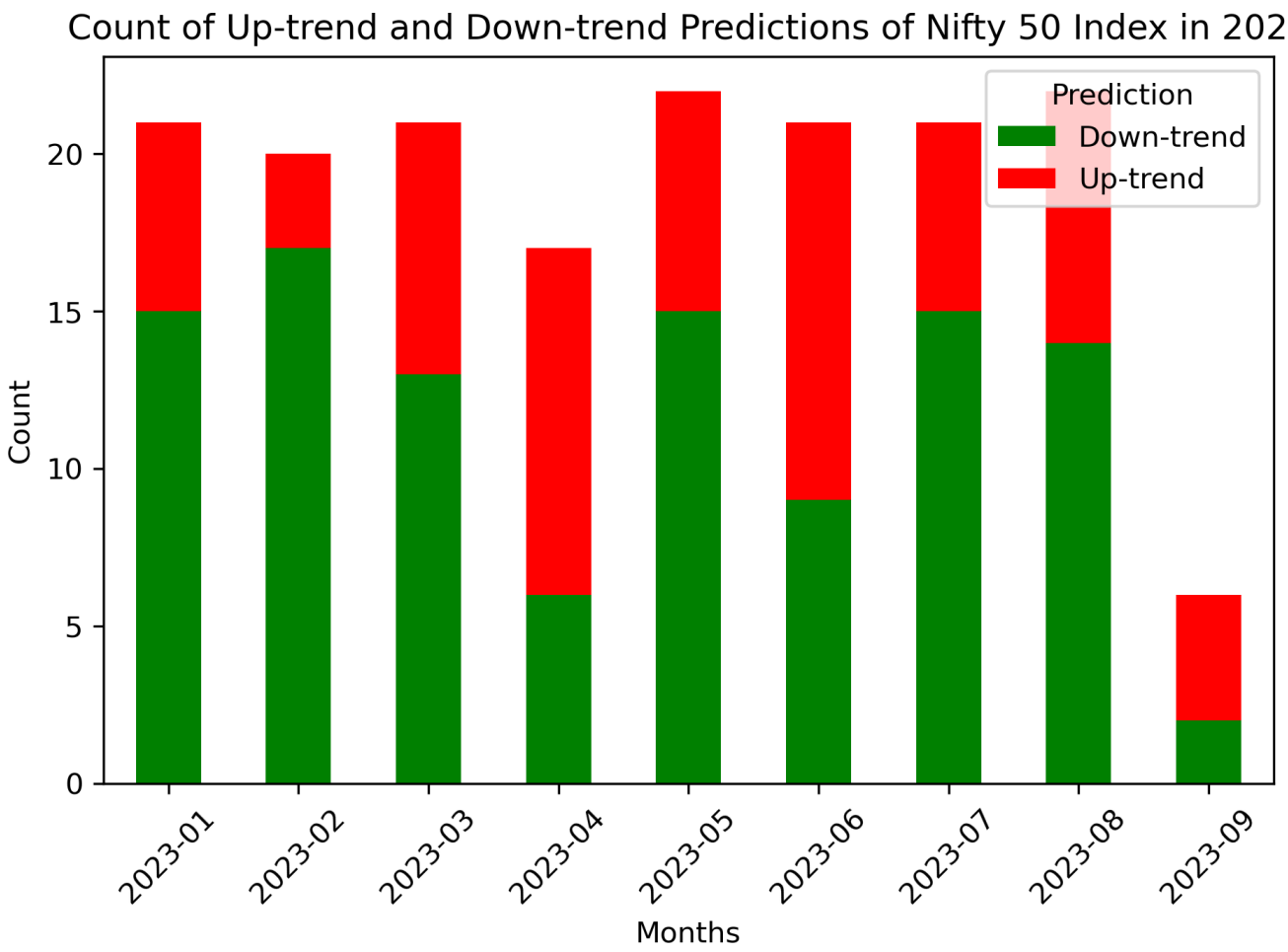

Figure 6 : Figure shows the prediction of up-trends and downtrends events in NIFTY 50 price. Blue colors depict the number of uptrends and red colors depict the number of downtrends in that particular month.

In figure 5 shows the total uptrend and downtrend events actually occurred in 2023. Blue color shows the down-trend events and red color up trend events. In Figure 6, we present a comprehensive overview of our analysis, shedding light on the stock market trends throughout the year 2023. The figure employs a color-coded visual representation to delineate the number of uptrends in a striking blue hue and down trends in a vivid red. This nuanced approach allows for a quick and intuitive grasp of the prevailing market sentiment.

These accuracies serve as a tangible measure of our model's performance, signifying its ability to forecast market movements with precision. Notably, our analysis yields accuracy rates that consistently surpass the 60% benchmark, a testament to the effectiveness of our model in capturing market dynamics on a month-to-month basis. The pinnacle of our achievements was witnessed in March, where our model achieved a remarkable 80% accuracy, indicative of its proficiency in forecasting stock trends during that period. On the other hand, February presents the lowest accuracy of 60%, underlining the subtle fluctuations and intricacies that can challenge even the most sophisticated predictive models.

Table 1 : Accuracy of our proposed model for different months of the year 2023.

Month Year	Down/Up Actual Count	Down/Up Prediction	Accuracy (%)
January 2023	13/8	15/6	71.42
February 2023	13/7	17/3	60.00
March 2023	13/8	13/8	80.95
April 2023	7/10	6/11	70.59
May 2023	11/11	15/7	63.64
June 2023	10/11	9/12	76.19
July 2023	12/9	15/6	76.31
2023 Average			70

These results not only underscore the robustness of our approach but also emphasize the significance of temporal nuances in the financial domain. This valuable insight guides us in further refining our predictive strategies, with the ultimate aim of achieving even higher accuracy rates and a more profound understanding of market dynamics.

5. Conclusion

In this study, we embarked on a comprehensive analysis of financial news data spanning the past two decades. Our goal was to construct a robust framework for predicting stock price trends. We meticulously curated our dataset by categorizing news articles to focus on the most pertinent content for our analysis. Prior to applying machine learning techniques, we employed rigorous text-cleaning processes tailored for natural language processing.

Our initial experiments featured the Bag of Words (BoW) model, coupled with a suite of machine learning algorithms, including K-Nearest Neighbors (KNN), Logistic Regression (LR), and Support Vector Machine (SVM). This initial phase yielded predictive accuracy spanning the range of 40

To bolster the predictive power of our model, we harnessed the capabilities of the BERT (Bidirectional Encoder Representation from Transformers) model. BERT facilitated the extraction of sentiment scores from each news article, encapsulating the underlying sentiment as either Bullish or Bearish. This sentiment-based approach was integrated with our machine learning framework, leading to a substantial leap in predictive accuracy.

The results of our study revealed that the incorporation of BERT-based sentiment analysis yielded accuracy rates in the range of 70%. This notable enhancement underscores the potential of advanced natural language processing and deep learning techniques in improving the precision of stock price trend predictions.

In conclusion, our research underscores the significance of sentiment analysis in financial markets. The integration of BERT-based sentiment scores with traditional machine learning models substantially improved our predictive accuracy. This study not only advances the understanding of market dynamics but also provides a practical framework for more accurate stock price trend predictions. As we venture further into the intricate world of financial analysis, the fusion of advanced language models with machine learning algorithms holds great promise for refining stock market predictions, opening new horizons for research and application in the financial sector.

The BERT-based sentiment analysis of financial news has shown remarkable potential in predicting trends in stock prices. However, this field presents a plethora of opportunities for further investigation and innovation. The following avenues for the future work offers exciting prospects for advancing our understanding and practical application of sentiment analysis in the financial domain.

Multimodal Sentiment Analysis The integration of multimodal data sources, including text, images, videos, and audio, can enrich our analysis of market sentiment. Future research can explore how combining multiple forms of media can provide a more comprehensive view of market sentiment, leading to more accurate predictions.

Event Detection Incorporating event detection into sentiment analysis models can provide a deeper understanding of how specific events impact market sentiment and stock prices. Research in this area can lead to more nuanced predictions.

Fine-tuning Strategies Exploring domain-specific pre-training or fine-tuning strategies for BERT models can enhance their performance in the financial domain. Customizing BERT for the intricacies of financial news can result in more accurate sentiment analysis.

Market Impact Assessment Exploring how different types of news and sentiment affect various categories of stocks can provide valuable insights. Understanding the differential impact of sentiment on technology stocks versus pharmaceutical stocks, for example, is an important aspect of future research.

Risk Assessment Sentiment analysis can be integrated into risk assessment models to predict market volatility and risk more accurately. Future work can focus on refining these models to make them more robust and reliable.

Acknowledgments

We would like to express our gratitude to Sovrenn.com for their generous support in funding this research, as well as for granting us access to their valuable resources

References

Aldhyani T. H., Alzahrani A., “Framework for predicting and modeling stock market prices based on deep learning algorithms”, *Electronics,*vol. 11, nº 19, p. 3149, 2022.

Bansal M., Goyal A., Choudhary A., “Stock market prediction with high accuracy using machine learning techniques”, *Procedia ComputerScience*, vol. 215, p. 247-265, 2022.

Devlin J., Chang M.-W., Lee K., Toutanova K., “BERT: Bidirectional Encoder Representations from Transformers”, *arXiv preprintarXiv:1810.04805*, 2018.

Joachims T., “Text categorization with support vector machines: Learning with many relevant features”, *European conference*

on machinelearning, p. 137-142, 1998.

Lin Y.-L., Lai C.-J., Pai P.-F., "Using deep learning techniques in forecasting stock markets by hybrid data with multilingual sentiment analysis",

Electronics, vol. 11, nº 21, p. 3513, 2022.

Pang B., Lee L., Vaithyanathan S., "Thumbs up? Sentiment classification using machine learning techniques", *Proceedings of the ACL-02conference on Empirical methods in natural language processing-Volume 10*p. 79-86, 2002.

Shah A., Gor M., Sagar M., Shah M., "A stock market trading framework based on deep learning architectures", *Multimedia Tools andApplications*, vol. 81, nº 10, p. 14153-14171, 2022.