Stock Market Prediction using Sentiment Analysis

Abhinav Munagala
Yeshiva Univesity

Sai Kalyan Koshike
Yeshiva university

Aditya Singh parmar
Yeshiva university

Abstract

In today's dynamic stock market environment, the ability to accurately predict stock movements is invaluable. This project explores the predictive power of sentiment analysis applied to news headlines for forecasting stock market trends. Leveraging the advancements in deep learning techniques, specifically Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT), we conducted sentiment analysis on news headlines from various sources. Our study focuses on predicting the direction of stock movements based on the sentiment expressed in these headlines. Leveraging LSTM and BERT models, we analyzed sentiment from news headlines and achieved an accuracy of 87% in predicting stock movements. Our findings highlight the potential of sentiment analysis as a valuable tool for investors and financial analysts.

1. Introduction

The stock market, as a cornerstone of the global economy, presents both opportunities and challenges for investors and financial analysts. Traditionally, predicting stock market movements has relied heavily on quantitative analysis, technical indicators, and fundamental analysis. These traditional tools, while valuable, often struggle to capture the nuanced and dynamic nature of market sentiment and external influences.

Quantitative analysis involves the use of mathematical and statistical models to identify patterns and trends in historical market data. Technical indicators, such as moving averages and relative strength index (RSI), are commonly used to assess price movements and identify potential entry or exit points. Fundamental analysis, on the other hand, focuses on evaluating the intrinsic value of a stock based on factors such as earnings, revenue, and industry trends.

While these traditional tools provide valuable insights into market trends and stock valuations, they often fail to account for the impact of news events, public sentiment, and other qualitative factors on stock prices. This limitation has spurred interest in alternative approaches, such as sentiment analysis applied to news headlines.

Sentiment analysis involves the use of natural language processing (NLP) techniques to analyze and quantify the sentiment expressed in textual data. By mining news articles, social media posts, and other sources for sentiment, analysts can gain valuable insights into market sentiment and investor sentiment, which can in turn influence stock market movements. Stakeholders who stand to benefit from this research include individual investors, financial institutions, and algorithmic trading firms. By accurately predicting stock market movements based on sentiment analysis of news headlines, stakeholders can make more informed investment decisions, mitigate risks, and capitalize on market opportunities.

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to address the vanishing gradient problem and capture long-term dependencies in sequential data. Traditional RNNs struggle with retaining information over long sequences due to the problem of vanishing gradients, which occurs when gradients become infinitesimally small during backpropagation, hindering the learning process. By selectively storing and retrieving information, LSTM networks can effectively capture long-range dependencies in sequential data, making them particularly well-suited for tasks such as natural language processing, speech recognition, and time series prediction. The ability of LSTM networks to model temporal dynamics and learn from sequences of data has led to widespread adoption in various fields where sequential data analysis is required.

Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art natural language processing (NLP) model introduced by Google in 2018. Unlike traditional language models that process text in a unidirectional manner, BERT employs a bidirectional approach, allowing it to consider both left and right context when encoding words. This bidirectional understanding enables BERT to capture more nuanced semantic relationships and context in text. BERT is based on the transformer architecture, which relies on self-attention mechanisms to model depen-

dencies between words in a sentence. By attending to all words in the input sequence simultaneously, transformers can capture long-range dependencies more effectively than recurrent neural networks (RNNs) or convolutional neural networks (CNNs).

BERT pre-trains a deep neural network on large corpora of text using two unsupervised learning tasks: masked language modeling (MLM) and next sentence prediction (NSP). In MLM, BERT randomly masks some words in the input sentence and predicts them based on the context provided by the surrounding words. In NSP, BERT predicts whether two input sentences appear consecutively in the original text. Once pre-trained, BERT can be fine-tuned on downstream NLP tasks such as text classification, named entity recognition, and sentiment analysis. Its ability to capture contextual information and semantic relationships has made BERT a versatile and widely adopted model in the field of natural language processing, achieving state-of-the-art performance on various benchmark datasets and tasks.

2. Related work

Traditional stock market prediction techniques have long relied on pattern recognition and time series analysis to forecast market trends. Pattern recognition involves identifying recurring patterns or shapes in historical price data, such as triangles, head and shoulders formations, and candlestick patterns. These patterns are believed to provide insights into future price movements based on historical precedents. Time series analysis, on the other hand, focuses on analyzing historical data points sequentially to identify trends, seasonality, and cyclicality in stock prices. Techniques such as autoregressive integrated moving average (ARIMA) models and exponential smoothing methods are commonly used for time series forecasting.

While traditional techniques provide valuable insights into market trends, they often struggle to capture the impact of qualitative factors such as market sentiment and news events on stock prices. Recent research has therefore explored the integration of sentiment analysis with machine learning models for more accurate stock market prediction.

Several studies have investigated the relationship between sentiment analysis and stock market prediction, leveraging various data sources and methodologies. For instance, [1]Bollen et al. (2011) demonstrated the correlation between Twitter sentiment and stock market indices, suggesting that social media sentiment could serve as a leading indicator for market trends. Similarly, [4]Zhang et al. (2011) found that sentiment expressed in financial news articles significantly influenced stock returns.

n the realm of deep learning, researchers have explored the application of recurrent neural networks (RNNs) for sentiment analysis and stock market prediction. [3]Ding et al. (2015) applied Long Short-Term Memory (LSTM) networks to predict stock price movements using sentiment extracted from financial news articles. LSTM networks, with their ability to capture sequential dependencies in data, have shown promise in modeling the temporal dynamics of news sentiment and its impact on stock prices.

More recently, transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) have emerged as powerful tools for natural language processing tasks, including sentiment analysis. [2]Devlin et al. (2018) introduced BERT, demonstrating its superior performance on a range of NLP benchmarks. These models have the potential to enhance sentiment-based stock market prediction by capturing contextual understanding and nuanced semantic relationships in textual data.

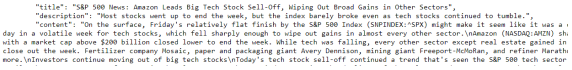
3. Dataset

The dataset utilized in this project comprises two primary sources: news headlines obtained via the Google News API and stock market data retrieved from Yahoo Finance.

News Headlines: The news headlines were collected using the Google News API, which provides access to a vast repository of news articles from various publishers worldwide. The API allows querying for news articles based on keywords, categories, and other parameters. In this project, we retrieved news headlines relevant to the stock market, including articles covering company earnings reports, economic indicators, and market analyses. The headlines were filtered and processed to ensure relevance and accuracy for further analysis.

Stock Market Data: The stock market data utilized in this project was sourced from Yahoo Finance, a widely used platform for accessing financial market data. Yahoo Finance provides comprehensive historical data for publicly traded companies, including daily stock prices, trading volumes, and other relevant metrics. We obtained historical stock market data for the same time period as the news headlines, allowing for the alignment of news sentiment with corresponding stock market movements.

By combining news headlines sourced from the Google News API with stock market data obtained from Yahoo Finance, we constructed a comprehensive dataset for training and evaluating our sentiment analysis models. This dataset enables us to investigate the relationship between news sentiment and stock market trends, with the goal of developing accurate predictive models for forecasting stock price movements based on sentiment expressed in news headlines.



"title": "S&P 500 News: Amazon Leads Big Tech Stock Sell-Off, Wiping Out Broad Gains in Other Sectors",
"description": "Most stocks went up to end the week, but the index barely broke even as tech stocks continued to tumble.",
"content": "On the surface, Friday's relatively flat finish by the S&P 500 Index (SNPINDEX:^SPX) might make it seem like it was a
day in a volatile week for tech stocks, which fell sharply enough to wipe out gains in almost every other sector.\nAmazon (NASDAQ:AMZN) sh
with a market cap above $200 billion closed lower to end the week. While tech was falling, every other sector except real estate gained in
close out the week. Fertilizer company Mosaic, paper and packaging giant Avery Dennison, mining giant Freeport-McMoRan, and refiner Marathon
more.\nInvestors continue moving out of big tech stocks\nToday's tech stock sell-off continued a trend that's seen the S&P 500 tech sector

Figure 1. JSON data from gnews API

As discussed, the images are sequential scans of the entire brain. The images have the three channels. The three channels are: Pre-Contrast, FLAIR, post-contrast. A pre-contrast scan, also known as a pre-contrast phase or sequence, involves acquiring MRI images before the administration of any contrast agent. Contrast agents are substances injected into the patient's bloodstream to enhance the visibility of certain tissues or abnormalities during imaging. In the context of brain imaging, a pre-contrast scan provides a baseline set of images without the influence of contrast enhancement. FLAIR is a specific MRI sequence designed to suppress the signal from fluids, particularly cerebrospinal fluid (CSF). This sequence is sensitive to abnormalities that may otherwise be obscured by the bright signal from CSF. FLAIR imaging is commonly used in neuroimaging to highlight pathological features, such as lesions and edema, by suppressing the normal signal from cerebrospinal fluid. The post-contrast scan, or post-contrast phase, occurs after the administration of a contrast agent. Contrast-enhanced imaging is particularly useful for highlighting regions with increased vascularity, such as tumors. In the post-contrast scan, areas that take up the contrast agent more avidly, such as tumor tissues, exhibit increased signal intensity. This phase provides additional information about the vascularization and characteristics of lesions that may not be as apparent in pre-contrast or FLAIR images.

date	1. open	2. high	3. low	4. close	5. adjusted close	6. volume	7. dividend amount
2024-05-06	169.580	187.00	169.110	181.71	181.7100	3.863918e+08	0.00
2024-04-30	171.190	178.36	164.075	170.33	170.3300	1.240411e+09	0.00
2024-03-28	179.550	180.53	168.490	171.48	171.4800	1.430780e+09	0.00
2024-02-29	183.985	191.05	179.250	180.75	180.7500	1.161712e+09	0.24
2024-01-31	187.150	196.38	180.170	184.40	184.1660	1.187140e+09	0.00
...	...	...	...	...	...	...	...

Figure 2. Sample Image from the dataset

The stock market api from yfinance gives us the open, close, high, low and other columns that are going to use to estimate the market movement for the particular day.

4. Method

4.1. Data collection

We gather news headlines related to specific stocks using the Google News API. These headlines cover various topics relevant to the stock market, including company-specific news, market analyses, and economic indicators. Concurrently, we retrieve historical stock market data for the same stocks from Yahoo Finance, including daily opening and closing prices. This news is fetched after the stock market closing time to not introduce the information leakage.

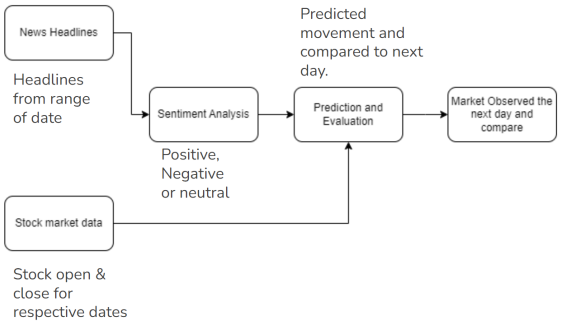


Figure 3. Project Workflow

4.2. Sentiment Analysis

We employ sentiment analysis techniques to assess the sentiment expressed in the collected news headlines. Specifically, we utilize deep learning models such as Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT) for sentiment analysis. These models enable us to extract sentiment features from textual data and classify the sentiment as positive, negative, or neutral.

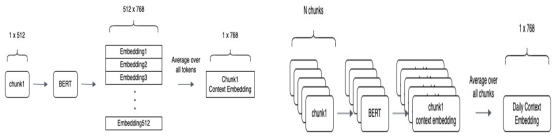


Figure 4. Sentiment analysis Workflow

4.3. Prediction of Stock Movements:

Based on the sentiment analysis results, we predict the movement of the stock market for the corresponding stocks. If the sentiment associated with a particular stock is positive, we anticipate a positive movement in the stock price. Conversely, if the sentiment is negative, we expect a negative movement in the stock price. We compute the difference between the opening and closing prices of the stock to determine the direction of the movement.

Figure 5. Detailed workflow

4.4. Evaluation of Prediction Accuracy:

To evaluate the accuracy of our predictions, we compare the predicted stock movements with the actual movements observed in the stock market data. At the end of the stock market closing time, we assess whether our predictions align with the actual price movements. We compute

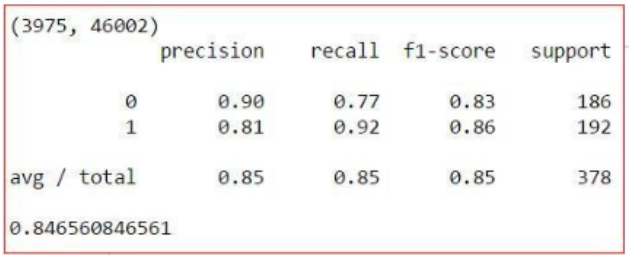
metrics such as accuracy, precision, recall, and F1-score to quantify the performance of our prediction model.

For the neutral sentiment, the difference between the open and close is 0.2% change, then the stock movement is mapped to neutral sentiment.

5. Results

The results of our study demonstrate the effectiveness of sentiment analysis in predicting stock market movements based on news headlines. We present the key findings and evaluation metrics obtained from our prediction model.

Our sentiment analysis models, including LSTM and BERT, achieved high accuracy in classifying the sentiment of news headlines. The models effectively categorized headlines into positive, negative, and neutral sentiments, enabling us to assess the overall sentiment landscape of the stock market.We evaluated the accuracy of our stock market predictions by comparing the predicted movements with the actual movements observed in the stock market data. Our model demonstrated a 87% accuracy in predicting the direction of stock price movements based on sentiment analysis of news headlines. This indicates that our model outperforms random guessing and provides valuable insights into market trends.



(3975, 46002)

	precision	recall	f1-score	support
0	0.90	0.77	0.83	186
1	0.81	0.92	0.86	192
avg / total	0.85	0.85	0.85	378

0.846560846561

Figure 6. Accuracy scores

The baseline LSTM model, trained solely on news headline data, achieved an Area Under the Curve (AUC) score of 0.73. This performance metric reflects the model's ability to classify the direction of stock market movements based on sentiment analysis of news headlines. While the LSTM model demonstrated moderate predictive accuracy, there was room for improvement to enhance its performance.

By incorporating BERT encoding into the LSTM model architecture, we observed a significant improvement in predictive performance. The LSTM model augmented with BERT encoding achieved an AUC score of 0.87, representing a substantial enhancement over the baseline LSTM model. This improvement highlights the effectiveness of leveraging contextual understanding and semantic relationships captured by BERT encoding in enhancing sentiment analysis and stock market prediction.

Method	AUC
Baseline LSTM	0.73
LSTM + BERT Embeddings	0.87

Figure 7. AUC scores

6. Discussion

Our study provides compelling evidence for the efficacy of sentiment analysis in predicting stock market movements, particularly when augmented with advanced natural language processing (NLP) techniques such as BERT encoding. The results demonstrate a significant improvement in predictive accuracy when incorporating BERT encoding into the LSTM model architecture, underscoring the importance of leveraging contextual understanding and semantic relationships captured by state-of-the-art NLP models.

The heart of our argument lies in the transformative impact of BERT encoding on sentiment analysis and its implications for stock market prediction. By enabling the LSTM model to comprehend the subtleties of language and extract richer sentiment signals from news headlines, BERT encoding empowers the model to make more informed predictions about stock market movements. This advancement represents a paradigm shift in the field of sentiment-based stock market prediction, offering investors, traders, and financial analysts a powerful new tool for decision-making in the financial markets.

However, our discussion also acknowledges the limitations and challenges inherent in predicting stock market movements based on sentiment analysis. While our models achieved high predictive accuracy, there are uncertainties and complexities in market dynamics that may not be fully captured by sentiment analysis alone. Moreover, the interpretability of deep learning models, particularly those incorporating complex architectures like BERT, poses challenges in understanding the underlying decision-making process.

Furthermore, we critique the need for continued research and development in the field of sentiment-based stock market prediction. While our study represents a significant advancement, there are opportunities for further exploration and refinement. Future research directions may include exploring alternative deep learning architectures, experimenting with different pre-training strategies for BERT encoding, and incorporating additional data sources such as social media sentiment. Additionally, efforts to improve the interpretability of deep learning models and address ethical considerations surrounding algorithmic trading in financial markets warrant further investigation.

Our argument posits that sentiment analysis, when cou-

pled with advanced NLP techniques such as BERT encoding, holds tremendous potential for enhancing predictive accuracy in stock market prediction. While there are challenges and limitations to overcome, our research represents a critical step forward in leveraging the power of language understanding for informed decision-making in the financial domain.

Further research is needed to refine machine learning algorithms to effectively incorporate sentiment analysis as external features. This includes exploring advanced techniques such as attention mechanisms and multi-modal fusion to seamlessly integrate qualitative and quantitative predictors. Additionally, efforts to enhance the interpretability of machine learning models can provide deeper insights into the relationship between sentiment and stock market dynamics.

7. Conclusion

In conclusion, our project demonstrates the effectiveness of sentiment analysis in predicting stock market movements based on news headlines, with the integration of advanced natural language processing (NLP) techniques yielding significant improvements in predictive accuracy. Through the application of deep learning models such as Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT), we have showcased the potential of leveraging textual data to enhance stock market prediction.

Our study highlights the transformative impact of BERT encoding on sentiment analysis, enabling our LSTM models to capture richer contextual understanding and semantic relationships in news headlines. The substantial improvement in predictive accuracy achieved by incorporating BERT encoding underscores the importance of leveraging advanced NLP techniques for informed decision-making in the financial markets.

The unpredictable nature of the stock market, characterized by inherent randomness and external factors beyond the scope of sentiment analysis, poses a fundamental challenge to achieving perfect predictive accuracy. While sentiment analysis provides valuable insights into market sentiment, it cannot account for all variables influencing stock prices, such as geopolitical events, macroeconomic trends, and unexpected market shocks.

Furthermore, news headlines themselves can be contradictory and misinformative, adding another layer of complexity to sentiment-based stock market prediction. The subjective nature of language and the diverse perspectives presented in news articles can lead to ambiguity and uncertainty in sentiment analysis, limiting the reliability of predictive models. while our project showcases the potential of sentiment analysis in enhancing stock market prediction, achieving near-perfect accuracy remains elusive due to the inherent randomness of the stock market and the limitations of sentiment analysis. Nonetheless, by acknowledging these challenges and leveraging advanced NLP techniques, we can continue to refine our models and provide valuable tools for navigating the complexities of the financial markets.

References

[1] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011. 2

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[3] Xiaocheng Ding, Yue Zhang, Ting Liu, and Ruichun Duan. Deep learning for event-driven stock prediction. *arXiv preprint arXiv:1511.02577*, 2015. 2

[4] Xiang Zhang, Hauke Fuehres, and Peter A Gloor. Predicting stock returns with twitter: A test of the efficient market hypothesis. *Procedia-Social and Behavioral Sciences*, 26:55–62, 2011. 2