Sentiment Analysis of Financial Markets Using Google BERT

Sri Vallabh Parupudi, Vignesh Chitra Ravikumar, Sai Charan Reddy Chennadi, Susruth Krishna Kagitala

The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State Univeristy, Pennsylvania, USA.

Contributing authors: sbp5860@psu.edu; vxc5173@psu.edu; sfc5929@psu.edu; sjk6753@psu.edu;

Abstract

Financial news has a big impact on stock performance in the fast-paced world of stock markets. It affects performance both technically and emotionally by influencing investor attitudes and views. Our study uses advanced Natural Language Processing (NLP) techniques, particularly the Bidirectional Encoder Representations from Transformers (BERT) model, to thoroughly analyse the sentiment expressed in these financial articles. Previous studies support BERT's effectiveness in sentiment analysis and indicate that hybrid models that combine BERT with architectures such as Bidirectional Gated Recurrent Unit (BiGRU) and Bidirectional Long Short-Term Memory (BiLSTM) improve classification accuracy even further Talaat, (2023). Motivated by these results, we go beyond bag-of-words models with our approach by using a pre-trained BERT framework that reduces bias and overfitting. Using this advanced methodology, our research aims to measure the emotional undertones of financial stories and the resulting influence on market behaviour, promoting a more strategic and knowledgeable financial decision-making process.

Keywords: Sentiment Analysis, Machine Learning in Finance, BERT, NLP

1 Introduction

In the field of finance, the stock market is known for its high volatility, drawing the attention of researchers aiming to understand and predict its fluctuations. Investors and market analysts meticulously analyse market behaviour to devise their investment strategies. Given the vast amounts of data generated daily by the stock market, it's challenging for an individual to process all available historical and current information to forecast future stock trends. Typically, there are two primary forecasting methods: Technical analysis and Fundamental analysis. Technical analysis relies on historical price and volume data to predict future trends, while Fundamental analysis involves scrutinizing a company's financial statements to derive insights about its health and potential future performance. However, the efficient-market hypothesis challenges both methods by suggesting that stock prices are inherently unpredictable due to market efficiency. This research adopts Fundamental analysis by leveraging news articles about a company to predict stock trends, classifying news as either positive or negative. The assumption is that positive news will likely lead to an increase in stock prices, whereas negative news could cause them to decline. The goal is to develop a model capable of predicting the sentiment of news articles and assessing its impact on stock prices. This study employs supervised machine learning for classification and other text mining techniques to determine news sentiment.

Bag of words approach was followed in the previous research by many researchers Schumaker et al., (2012) has used the AZFintext is designed as a financial news analysis tool with a focus on predicting stock market prices based on the information extracted from real-time financial news articles. Two distinct methods for natural language processing in the finance industry are AZFinText and BERT. AZFinText was created in the early 2000s with a specific focus on predicting stock market patterns through the analysis of financial news. It provides quick and specialised insights but has limited versatility because it depends on conventional statistical methods and rule-based systems designed for financial data. Conversely, Google's more recent invention, BERT, uses a deep learning framework that is built around the transformer architecture. As a result, it can comprehend contextual subtleties in a variety of NLP tasks. By fine-tuning on domain-specific datasets, BERT's versatility and cutting-edge performance make it appropriate for a range of applications, including financial sentiment analysis. Although BERT needs more processing power, its adaptability and breadth of linguistic comprehension are superior to those of programmes such as AZFinText.

2 Literature Review

Schumaker et al., (2012) developed the AZFinText system to assess how financial news' objectivity/-subjectivity and sentiment affect stock predictions. Analyzing 2802 S&P 500 articles from 2005, they used OpinionFinder and support vector regression to determine sentiment impacts. The results demonstrated that subjective, negatively-toned articles outperformed the baseline model in directional accuracy (59%) and returns (3.3%). The findings suggest that investor behavior may be influenced by sentiment, particularly in market downturns, highlighting the value of incorporating sentiment analysis in financial predictions.

Li et al., (2014) investigated the role of sentiment analysis in forecasting stock returns, comparing it to traditional bag-of-words models. Utilizing news from the Hong Kong Stock Exchange and employing sentiment dictionaries like Harvard IV-4 and Loughran-McDonald, their models significantly outperformed others, especially at individual stock and sector levels. The results confirmed the effectiveness of sentiment analysis in real-time trading scenarios, advocating for further enhancement of sentiment dictionaries and the exploration of intra-day trading strategies.

Yadav et al., (2019) presented a method for conducting sentiment analysis on financial news to identify top investment opportunities. They created a labeled dataset correlating news with market trends using metrics like net buying pressure and constructed feature vectors from news headlines through natural language processing techniques. Employing classifiers like SVM and naive Bayes, their structured approach showcased the potential of sentiment analysis in uncovering significant stock market insights.

Kalyani et al., (2016) explored how financial news sentiment correlates with stock price movements using machine learning. Despite the Efficient Market Hypothesis suggesting unpredictability, they collected and processed data on Apple Inc., using sentiment analysis to predict stock trends. Their models, especially Random Forest and SVM, demonstrated high accuracy in detecting stock movements based on news sentiment, validating the predictive power of news sentiment analysis in stock market investments.

Baker and Wurgler, (2007) developed a macroeconomic sentiment index to study its impact on stock market returns, using proxies like NYSE turnover and IPO volume. Their analysis indicated a clear relationship between sentiment shifts and overall market performance. The study suggested that investor mood significantly affects market dynamics, particularly influencing the performance of speculative versus bond-like stocks, depending on the prevailing sentiment.

Van Bunningen et al., (2004) proposed a method to analyze and predict stock prices using NLP on news articles, focusing on the pharmaceutical sector. By extracting features like favorability from news content and using SVM for analysis, their model outperformed basic predictions, illustrating the potential benefits of applying NLP to understand and anticipate stock price movements based on news content.

3 Methodology (Model Implementation)

In this study, we plan to focus on implementing advanced NLP techniques to implement sentiment analysis. For that we have employed BERT (Bidirectional Encoder Representations from Transformers) model, to analyse sentiment in financial news. Understanding the sentiment of the market plays a crucial role in taking financial decisions and the algorithm that we have used can definitely help in understanding the sentiment of financial news from different sources. To collect the financial news data we have used the help of News API. By taking the financial news from News API and analysing the sentiment using BERT we plan to find if the actual share market performance is correlated with the sentiment of the news.

3.1 System Design

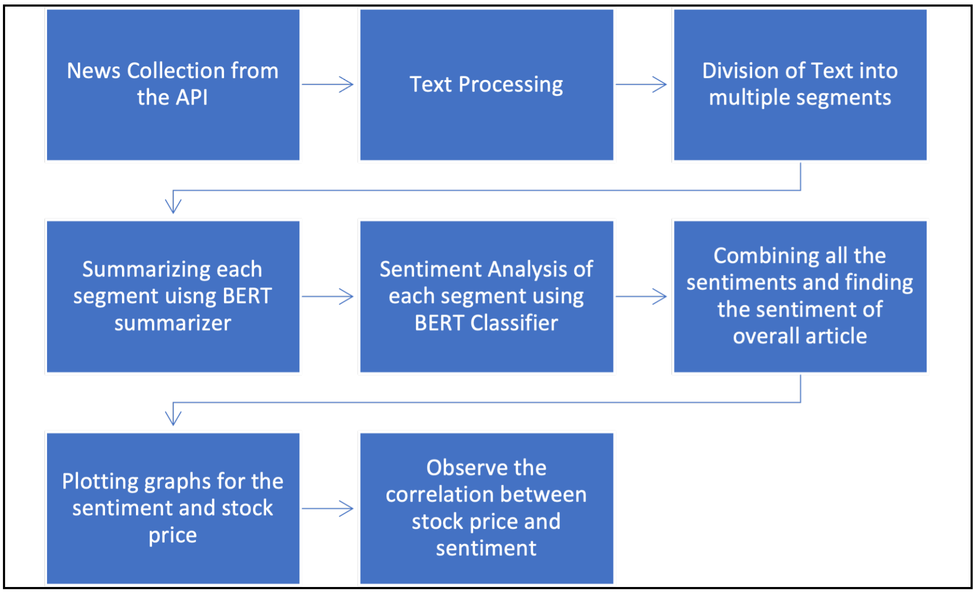


Figure 1: System Design

3.2 Sentiment Analysis

Sentiment analysis is the methodology that is used to detect the emotion of the digital text and let us know whether it is positive or negative or neutral. All over the world many companies implement this process in their systems to analyse different types of data such as customer emails, customer support chatbots, community blogs and reviews. Nowadays it is adversely used by social media applications, and a lot of research is done in developing better algorithms for more accurate sentiment detection.

Two primary approaches are usually used in the creation of sentiment dictionaries for word evaluation: semi-automatic and manual building.Li et al., (2014) When constructing a dictionary semi-automatically, a manual selection of seed words is made first. The basis is laid by these seed words. The program then applies predetermined rules or algorithms to further extend the lexicon. Conversely, manual construction entails having linguistic specialists create the dictionary from scratch. When compared to the semi-automatic method, this strategy typically yields a smaller dictionary but has higher sentiment categorization accuracy.

3.3 Working of BERT Model

BERT (Bidirectional Encoder Representations from Transformers) is a powerful and popular language representation model developed by Google AI in 2018. BERT has significantly advanced NLP tasks by using Transformer architecture. The main advantage of this model is it is pre-trained on vast amount of text data. Ravichandiran, (2021)

Pre-Training BERT:

BERT is pre-trained with the help of two main unsupervised learning tasks

1. Masked Language Model (MLM): During pre-training, BERT randomly masks some of the words in a sentence and then tries to predict the masked words based on the context provided by the surrounding words. Ravichandiran, (2021)
2. Next Sentence Prediction (NSP): BERT is trained to predict whether a sentence logically follows another sentence in a given text corpus. This will help BERT to learn to capture relationships between sentences and understand the overall context of longer sentences forming a paragraph. Ravichandiran, (2021)

After pre-training, BERT can be fine-tuned for specific downstream tasks and in our case we have used BERT in two different cases

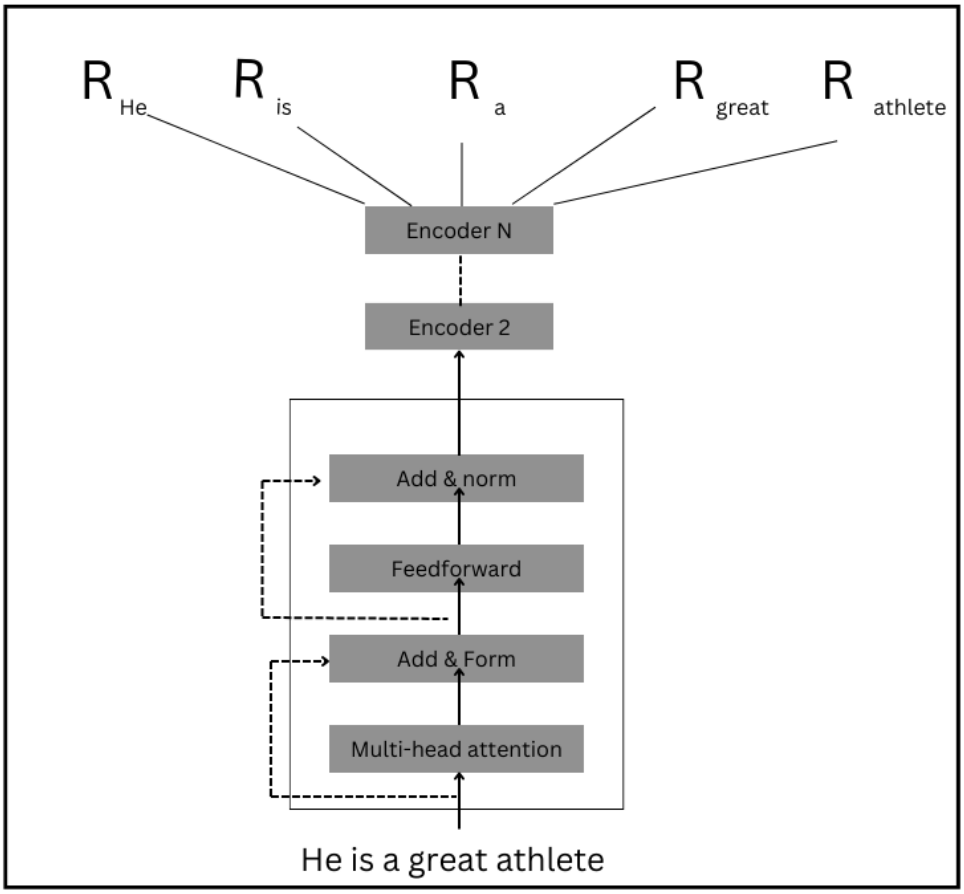
1. BERT Summarizer
2. BERT Sentiment Detection

Major Advantages of BERT:

1. BERT captures rich contextual information of words.
2. BERT is a pre-trained model. So, there is no hassle of training the model with a dataset and a dataset which is pre-processed can be directly used for the prediction.
3. BERT considers context from both directions (left-to-right and right-to-left), leading to better understanding of word relationships.

3.3.1 Example of working of BERT

1. Here is an example of how BERT encodes. Let us take a sentence " He is a great athlete".
2. Next, we input the sentence into the transformer's encoder and get the contextual representation (embedding) for each word in the sentence as the output.
3. After we feed the sentence the Multi-head attention process understands each word in the sentence and relates each word in the sentence to all the words in the sentence to learn the relationship and understand the context.
4. The above is the process of how a single encoder works, in a similar way we can stack upto N number of encoders and follow the same process.



;

Figure 2: Generation of each word in a sentence using BERT Ravichandiran, (2021)

3.4 Data Collection: Description of Stock Market Data Used

The data collection for our research involved utilizing News API, a versatile open-source API that aggregates news articles from a wide range of sources. For the purposes of our study. We purposely restricted the search to US-centric sources. This choice is in line with our study goal of analysing patterns and trends in companies listed on the National Association of Securities Dealers Automated Quotations (NASDAQ), which makes sure that the news is relevant to the particular market environment we are looking at.

Furthermore, in addition to our textual news data research, we utilised the Trading View platform for the stock price information. Trading View is renowned for offering a wide range of tools to meet different needs in financial analysis. But we especially used its line graph functionality for our investigation. Since our goal is to track stock performance and discover trend patterns, line graphs were selected due to their ability to clearly and effectively illustrate price movements over time.

3.5 Data Pre-Processing Steps: Data Cleaning and Formatting

For data pre-processing we have followed a step-by-step process and it goes as follows

3.5.1 Data Collection

The process begins with the use of the function 'getNews', which interfaces with the News API to gather news articles. It uses specific queries related to the stock market and financial sectors to filter relevant news articles. The function ensures that the article's title name has the company name and to ensure that we have given the company name as well as ticker symbol so that when we call the API it picks up all the news articles that have the financial or stock market related news. We have ensure that we put them in the date for exactly 10 days as we plan to analyse the market for the 10-day period. For instant news, the method retrieves headlines from the top headlines endpoint; for a comprehensive collection of articles, it then switches to the 'everything' endpoint depending on a desired language and date range.

3.5.2 Data Pre-Processing

De-Contraction
This step is used to normalize text, making it easier to expand English contractions. For example "won't" will be changed to "will not". The function used is `decontracted(phrase)`

Text Cleaning
With the help of `clean_text(text)` function we have removed HTML entities, hyperlinks, hashtags, and Twitter-related artifacts such as "RT" for retweets. It also helps in limiting the repetition of any character within a word to a maximum of two consecutive occurrences, which can help in normalize words that have been elongated for emphasis (e.g., "sooo" becomes "so").

Chunking Text
This is a very important function and for this we have used `chunk_text(text, max_length=512)`. This splits the whole articles into chunks of 512 word length. This is important because the BERT algorithm can only take up to 512 words during summarization at once.

Article Summarization
After chunking with the help of summarization module available in BERT we have summarized each chunk which is further sent to full article summary retrieval.

Full Article Content Retrieval
In this step the summary of each chunk is combined and sent into the pipeline further to perform the sentiment analysis using BERT.

Procedure

1. A fine-tuned pre-trained BERT model is used for sentiment classification and the sentiment of each summarized article is analysed using a sentiment analysis pipeline.
2. The pipeline is used to process the text, and a sentiment label (e.g., positive, negative, neutral) is obtained for each chunk of text. The labels from all chunks are then aggregated to determine the overall sentiment of the article.
3. To give the article a final sentiment label, a majority vote among all chunk sentiments is taken. This label is a representation of the article's main idea.
4. All the processed data, including summaries and sentiment labels, are compiled into a structured format, such as a pandas DataFrame, for further analysis and storage.
5. All the outputs are then written into a CSV file for further analysis.

4 Sentiment Analysis Results

With the help of Tableau we have plotted our results of each stock. The graphs represented show two plots in the same chart. The bar chart represents the number of negative and positive articles on each day and the line represents the performance of the stock during the designated time (2024-03-24 to 2024-04-03).

Even though we cannot take a direct correlation as correlation is not the reason for causation but still the plots help us analyse whether the decrease or increase in stock price has aligned with the number of negative or positive news that we have taken in our dataset. This pattern can be justified by the fact that most of the news is discussed on the prediction done by the analyst.

Amazon's Stock

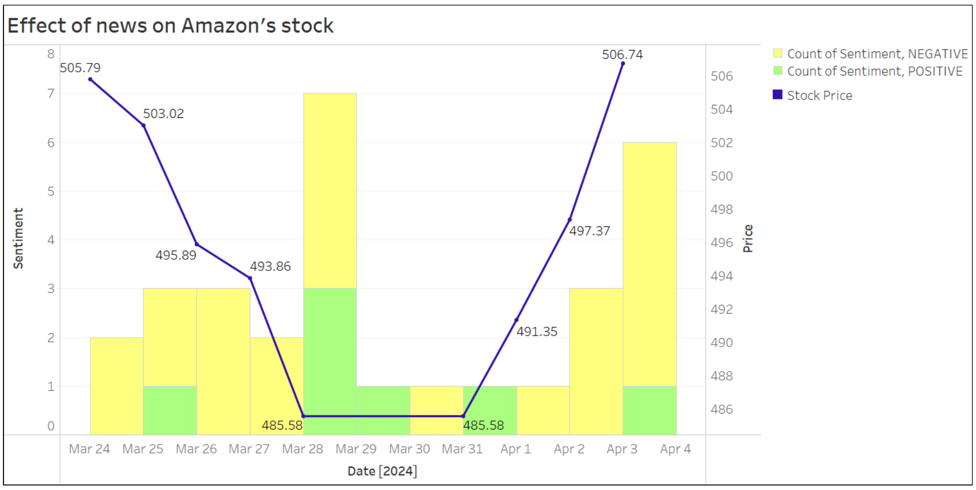


Figure 3: Effect of news on Amazon's stock

- Negative news sentiment appears to have some correlation with stock price movement. When negative sentiment peaked on March 26, the stock price was at a lower point. As negative sentiment counts lowered towards March 31, the stock price began to rise significantly on April 1 and continued to climb to a high on April 4.
- Positive news sentiment remained relatively stable and low compared to negative sentiment, which suggests that negative sentiment had a more discernible impact on stock price fluctuations during this period.

Google's Stock

- The stock price did not appear to react strongly to negative news sentiment, which varied considerably without a clear corresponding movement in stock price.
- A significant spike in positive sentiment on March 27 didn't coincide with a large increase in stock price, suggesting that for this period, Google's stock price might have been influenced by other factors beyond the analysed news sentiment.

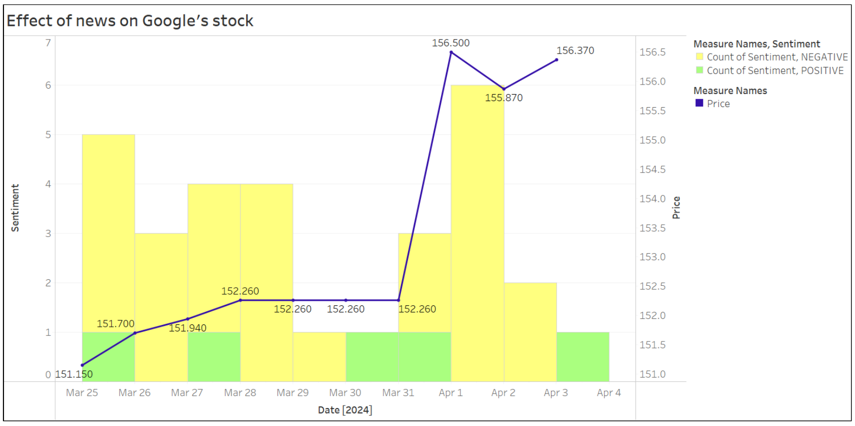


Figure 4: Effect of news on Google's stock

META's Stock

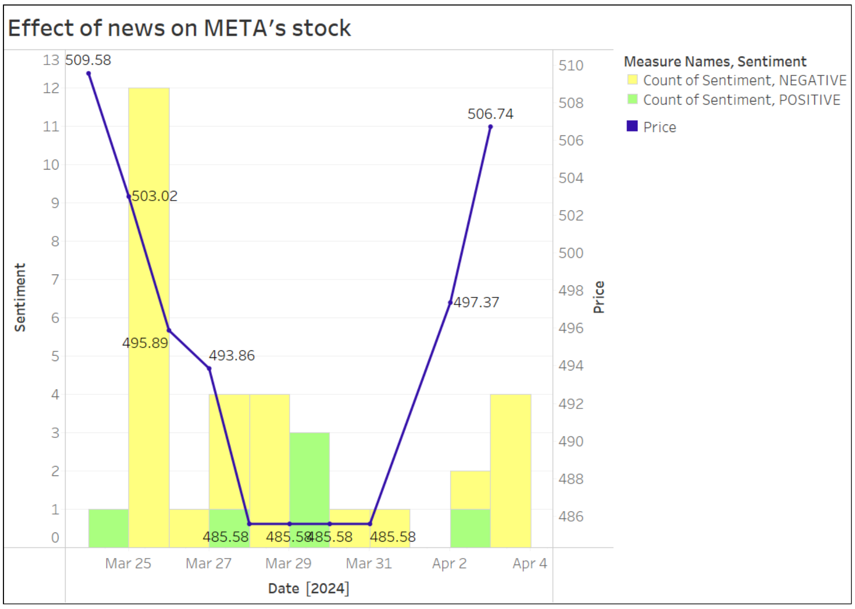


Figure 5: Effect of news on META's stock

- Similar to Amazon, there seems to be a connection between the news sentiment and stock price for META. The stock price dropped significantly as negative sentiment reached its highest level on March 25. As negative sentiment decreased after March 31, the stock price showed a marked recovery.
- Positive sentiment was again lower in comparison to negative sentiment but did show a slight increase on the same days as the stock price rose, suggesting a potential relationship between positive news sentiment and stock performance.

Microsoft's Stock

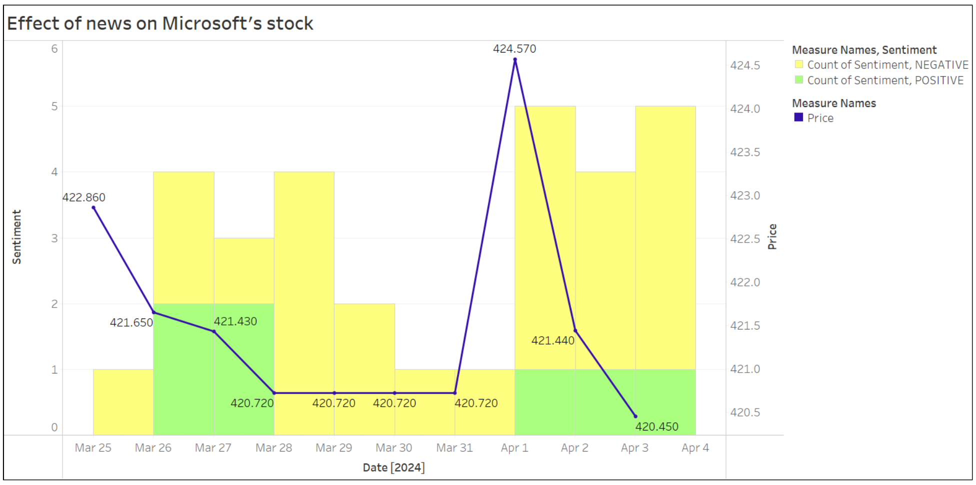


Figure 6: Effect of news on Microsoft's stock

- Microsoft's stock price showed an interesting pattern with a steep drop on April 2, just as negative sentiment spiked. This could indicate a possible immediate reaction to negative news in the market.
- Positive sentiment counts were lower and did not show a clear correlation with the stock price movements, similar to the trend observed with Google's stock.

NVIDIA's Stock

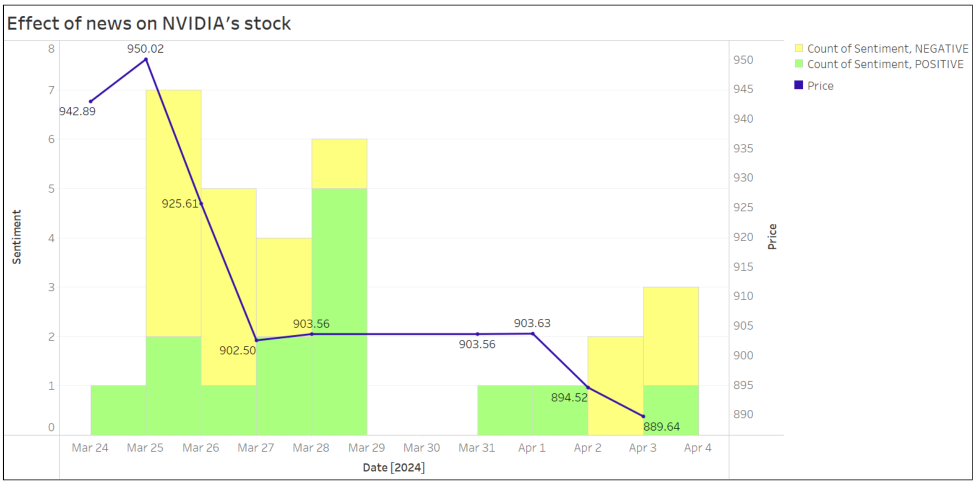


Figure 7: Effect of news on NVIDIA's stock

- NVIDIA's stock showed a strong decrease in price on April 2, coinciding with the highest negative sentiment count. Before this, the negative sentiment counts were rising while the stock price was falling, which suggests a negative correlation between stock performance and negative news sentiment.
- The positive news sentiment count was low and didn't show any significant changes to correlate closely with the stock price movements, which were more dramatically aligned with the negative sentiment spikes.

5 Future Research

We intend to further this research going forward, aiming for a more profound exploration of this topic and the application of enhanced methodologies and data quality improvements. Here is our intended approach:

1. Better and Bigger Dataset: We plan to take a bigger dataset by the use of another API as our present API has limited us to only take 30 articles per company. So, by taking a bigger dataset we have more data to do our analysis.
2. Integration of Real-Time Data: We plan to also include the real-time data of the stock market with the help of machine learning techniques. This incorporation of real-time analytics could allow for a more immediate understanding of sentiment shifts.
3. Cross Referencing Economic Indicators: Integrating economic indicators with sentiment analysis could provide a more holistic view of the factors influencing stock prices. This includes examining how news sentiment interacts with indicators like interest rates, inflation, employment data, and GDP growth.

6 Conclusion

Our in-depth study employed NLP tools, with a particular focus on the BERT model, to quantitatively assess the sentiment conveyed in financial news articles. Our findings show a significant correlation between negative news sentiment and the immediate performance of stock prices of Amazon, Google, META, Microsoft, and NVIDIA. The negative sentiment, captured through our sentiment analysis, frequently preceded notable declines in the stock market, suggesting that investor sentiment, as reflected in news coverage, can be a potent indicator of market trends.

In contrast , the impact of positive news sentiment on stock prices was less evident, indicating the potential dominance of other factors in driving market optimism. This suggests that while negative news can have a slight effect on stock prices, the positive news is insufficient on its own to support upward trends indicating that there are very less number of cases where the stock price increased because of positive news. This might be because the fetching of the news articles resulted most of the news to be negative, may be taking bigger data can help in better prediction.

References

1. Talaat, A. S. (2023). Sentiment analysis classification system using hybrid BERT models. Journal of Big Data, 10(1), 110.
2. Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. Decision Support Systems, 53(3), 458-464.
3. Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. Knowledge-Based Systems, 69, 14-23.
4. Yadav, R., Kumar, A. V., & Kumar, A. (2019). News-based supervised sentiment analysis for prediction of futures buying behaviour. IIMB Management Review, 31(2), 157-166.
5. Kalyani, J., Bharathi, P., & Jyothi, P. (2016). Stock trend prediction using news sentiment analysis. arXiv preprint arXiv:1607.01958.

6. Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. Journal of economic perspectives, 21(2), 129-151.
7. van Bunningen, A. H., Nijholt, A., Poel, M., & van Otterlo, M. (2004). Augmented trading. Master of Science Thesis.
8. Ravichandiran, S. (2021). Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT. Packt Publishing Ltd.

Appendix

Source Code:
https://drive.google.com/drive/folders/1hzmhG-L4MqSGeaY3YPpWnXtmD8ELh--P?usp=sharing