

# Sentiment Analysis of Stock Market: Extracting and Deciphering Market Sentiments from Digital Seas

Emran Khan<sup>1</sup> and Praveen Ailawalia<sup>2</sup>

<sup>1,2</sup>Department of Mathematics, University Institute of Sciences, Chandigarh University, Gharuan- Mohali, Punjab, INDIA  
Email: emrann.khan03@gmail.com, Praveen\_2117@rediffmail.com

**Abstract—** In order to evaluate people's behavior and activities, information collecting has become essential. Online learning is done through sharing and exchanging ideas on the Internet. On social networking sites like Twitter, Facebook, Google+, Yahoo, and MSN, opinions are shared. Over the past 20 years, the development of Internet-based technology has led to considerable modifications to the Indian stock market. Many data mining approaches are used to manage changes in the SM. It is considered that financial news headlines have an impact on stock trend prices returning. For managing predictions related to SM, machine learning may provide an approach that is more dependable and precise. Opinion mining, also known as sentiment analysis, is a method for locating and obtaining subjective data from text by examining the opinions, assessments, sentiments, attitudes, and emotions of users. In this study, it was important to use sentiment analysis while predicting stock price using stock market indicators like Tesla, BTC, and GME. Finally, we offer some recommendations for future research and make some conclusions.

**Index Terms—** Stock market, Sentiment analysis, emotion, social media, accuracy, NLP, stock market prediction, opinion mining, machine learning.

## I. INTRODUCTION

People frequently consider other people's opinions while making decisions, and this was especially true before the World Wide Web. Guidance from friends, acquaintances, and coworkers were highly important while making decisions on a daily basis. However, more people are using the Internet to express their thoughts to others. Users of social networks and social media can express their thoughts about a wide range of subjects, including current affairs, entertainment, news, and products [1]. The study of classifying opinions and expressions as positive, negative, or neutral is known as sentiment analysis also defined as SA. The most straightforward way to define SA is as an analysis carried out to extract data based on user sentiment, while there are alternative interpretations as well. Opinion mining, often referred to as sentiment and emotion analysis, is the study of beliefs, ideas, circumstances, emotions, and actions in text [2]. Financial market emotions may now be detected by an entire market [3]. For investors and scholars, stock forecasting is essential. There have always been both short-term and long-term changes in stock values. Daily news reports are crucial for informing the public about corporate or financial matters and for exposing SM trading techniques. They also have a key impact in forecasting stock prices. This study's main objective is to anticipate SM movement using news articles. Typically, news articles regarding a particular industry discuss the company's operations and future prospects for its stock. It is feasible to utilize

financial data in fundamental research to anticipate stock price movements at a much quicker pace when more financial data becomes accessible [4-7]. Figure 1 illustrates how the aforementioned SM indicators are impacted by news from the financial markets and social networking sites.

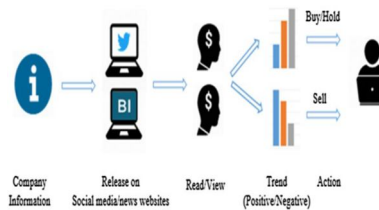


Figure 1: A thorough investigation on the effects of financial news and social media on stock market developments

The structure of the paper is as follows. Section II includes prior research as well as a review of the literature. Techniques for classifying emotions are presented in Section III. In Section IV, the recommended work with implementation is provided. Section V concludes the paper.

### A. Objective of research

- The goal of our research is to develop a framework for evaluating the stock emotion by using various online platforms.
- We provide a helpful market movement analysis by outlining the relationship between and application of financial news and stock indicators.

## II. LITERATURE REVIEW

The literature has investigated a number of strategies for forecasting stock market behavior and price trends. Some of this research focuses on utilizing sentiment analysis of news or tweets in addition to stock prices to increase the precision of predictions. Some, like [8], focus on price forecasting over a variety of time frames. Furthermore, other study approaches, including [9], showed a strong correlation among stock price fluctuations and financial news. Finally, research investigations like [10], [11] were carried out to increase prediction accuracy.

### A. Studies Invoking Information Analysis from Social Media

Utilizing information collected from tweets, some researchers created a method that anticipate stock price swings with a precision of up to 75.12% [12]. Bing used a method to look at developments in daily stock prices and open tweets. To find patterns in the correlations among investor sentiment and stock prices, data mining and NLP methods have been coupled. This research investigates internal linkage in multilayer hierarchical frameworks and shows how it relates to the top layer of unstructured data and the inner levels. For historical stock values, this approach only considers typical closing prices. Some researchers proposed a technique that uses sentiment analysis of tweets to forecast the performance of the Indonesian stock market [13].

### B. Research Based On News Analysis

Only a few studies [14] looked at how financial news impacted predictions of stock market values and daily changes in index movements. In order to improve prediction accuracy, this effort seeks to integrate technical analysis with a rule-based classifier. The forecast algorithm takes into account both the financial news and a regular basis average monthly stock price. Studies have suggested a mechanism for predicting changes in the market by leveraging non-quantifiable information found in news stories [15]. Senti-wordnet 0.3 and the statistical parameter-based module are used in the construction of the NLP approach in this model. The model generated polarization and activity as either positive or negative by using the open and closing intrinsic values of the stock.

### C. Research Gap

The preceding studies' stock trend forecast accuracy ranged from 75% to 80% and was based on textual data analysis. The range of forecast accuracy for news polarities is limited. The study presented in this paper aims to minimize losses by producing extremely precise forecasts based on stock and sentiment data.

## III. METHODOLOGY

The two main categories of algorithms for classifying sentiment are machine learning techniques and lexicon-based methods. These techniques are explained as follows:

#### *A. Approach using Machine Learning*

Sentiment analysis employs both supervised and unsupervised learning methodologies as machine learning algorithms.

#### *B. Unsupervised Learning*

Unsupervised learning is learning by observation with no clear desired output linked to the input. The idea is to let the computer learn on its own without being explicitly instructed. Clustering is a well-known unsupervised learning technique that identifies similarities between items in the training data. Among the metrics used to define the cluster similarity parameter is Euclidean distance. Several clustering algorithms include Hierarchical, K-means, Hidden Markov Models, Self-Organizing Maps and Gaussian Mixture Models [16].

#### *C. Supervised Learning*

A known dataset is utilized in supervised learning to predict the result. A training set and a test set of documents are needed for supervised learning. Training sets are used to learn about the many characteristics of documents, while test sets are used to assess the effectiveness of the classifier. The following supervised learning methods are as follow:

- Classification using decision trees
- Rule-based classifier
- Probabilistic classifiers
- Linear classifier

#### *D. Lexicon-based strategy*

Lexicon-based approaches for sentiment analysis are reliable and produce strong cross-domain results. This approach is predicated on the idea that contextual sentiment orientation is the total of the sentiment orientations of all the words. This tactic is further split into two categories as follows:

- Dictionary-based methodology
- Corpus-based strategy

#### *E. Data Collection*

The procedure for gathering data, its sources, and its organization are all covered in this subsection. We have collected data from the specific news source which was highly reputed and diverse. Stock news was obtained from Yahoo! Finance. We analysis the financial news sources for the stock market prediction. We choose online news platform to collect the news sentiments on the basis of real time data.

#### *F. Preprocessing*

To prepare the news headlines for the machine learning algorithms, the following procedures are followed.

- Word tokens are created from news documents.
- HTML and other tags are eliminated, including cashtags and the author tag (@). These tags must be deleted because machine learning algorithms cannot identify attitudes using them since they contain no meaningful information.
- Irrelevant URLs are eliminated.
- Additionally, stop words are removed. These are the terms (such is, the, are, an, etc.) that are commonly seen in news headlines but don't provide classifiers with any useful information.
- The act of changing words into their similar stems is known as stemming.
- Duplicate and irrelevant news articles are eliminated.

#### *G. Feature Extraction*

A characteristic that is developed in news data files is called News Sentiments. These sentiments are the combined feelings of specific news stories based on real-time data.

### **IV. RESULTS AND DISCUSSION**

#### *A. Implementation Outcomes*

The main goal of the suggested model is to extract real-time server data in Python so that sentiment analysis may be done on the sets of data that were obtained from the online news site. In this instance, the Windows 13 terminal

command-line environment is used to install necessary applications, such Beautiful Soup, once Python has initially been enabled on the machine. Software for Beautiful Soup may be installed on Windows 13 platforms by using system package management. The tool Beautiful Soup enables users grab content from a specified webpage and preserves the necessary data by using the Python library to get live data. The Python standard library's HTML parser is supported by Beautiful Soup. For demonstration purposes, we obtain current Tesla, BTC, and GME data that can be used for sentiment research in the future. Following modules and library are used for implementation:

- *Transformer*

A Python library called Transformers offers a variety of tools and pre-trained models for activities related to natural language processing (NLP), such as text categorization and language synthesis.

- *Requests*

A Python library called Requests is used to submit HTTP requests. It enables data retrieval from online servers and interaction with web services.

- *Beautiful Soup*

It is a python library for web scrapping. It is used for parsing HTML and XML documents making it easier to extract and manipulate data from web pages.

- *Pegasus Tokenizer*

The Hugging Face Transformers library includes the Pegasus tokenizer, which is a tool especially for tokenizing text while interacting with the Pegasus model. Text data preparation for Pegasus model inputs is done using it.

- *Pegasus for conditional generation*

It is also a component of the Hugging Face Transformer library, which is an effective toolkit for handling different NLP tasks and models. Additionally, it offers a pre-trained Pegasus model for jobs involving conditional text production, including text summarization.

- *Sentence piece*

It is a python library used for text tokenization. Its commonly used in NLP tasks to split text into smaller units such as sub word tokens, making it suitable for various text processing tasks.

- *Sorting of URL*

The first sort key is used to sort the records; ties are then broken by the second sort key, which breaks ties by the third sort key, and so on. In python sorting the URLs in the list are done according to their top-level domain is the work at hand. A top-level domain referred to TLD is a domain that is located at the top of the Internet's hierarchical Domain Name System. Like org, com, and edu. This is primarily employed when it is necessary to scrape the sites and arrange the URLs by top-level domain.

```
[ 'GME': ['https://finance.yahoo.com/news/gmeston-reports-fourth-quarter-fiscal-2020-08-26.html',
'https://www.fastcompany.com/90869514/gmeston-gme-stock-price-going-up-profit-retail-sdy',
'https://finance.yahoo.com/news/eyes-fed-decision-key-stocks-152603931.html',
'https://finance.yahoo.com/news/existing-home-sales-nike-earnings-162923739.html',
'https://www.google.com/search?q=30yahoofinance&btnG=30news&oeq=301',
'TSLA': ['https://ca.finance.yahoo.com/news/tesla-wireless-charging-platform-hands-on-well-made-and-sustainable-1305575.html',
'https://finance.yahoo.com/news/analyst-downgrades-tesla-stock-after-98-pow-13016397.html',
'https://finance.yahoo.com/news/tesla-investor-day-three-things-to-watch-17070139.html',
'https://ca.finance.yahoo.com/news/tesla-debt-no-longer-junk-36474581.html',
'https://www.google.com/search?q=30yahoofinance&btnG=30news&oeq=301',
'BTC': ['https://finance.yahoo.com/news/arks-cathie-wood-bitcoins-strength-32508695.html',
'https://finance.yahoo.com/news/bitcoin-surged-above-30000-11-it-a-safe-haven-from-banking-152435558.html',
'https://finance.yahoo.com/news/bitcoin-moves-above-us-26-82561450.html',
...
'https://finance.yahoo.com/news/bitcoin-set-extend-70-surge-864212977.html',
'https://www.google.com/search?q=30yahoofinance&btnG=30news&oeq=301',
```

Figure 3: Sorting of URL

The following stages are listed in the flowchart as being necessary for sentiment analysis utilizing Python scripting language:

The proposed model downloaded the current emotions of the Tesla, BTC, and GME data from multiple URLs like finance.yahoo.com, fastcompany.com, and msn.com for implementation purposes. Then used a Python script to fetch the data are then preprocessed so that NLP models can be created from them. The goal is to combine sentiment scores and stock emotion predictions into a single dataset in order to increase the likelihood that Tesla, BTC, and GME stock emotion can be predicted.

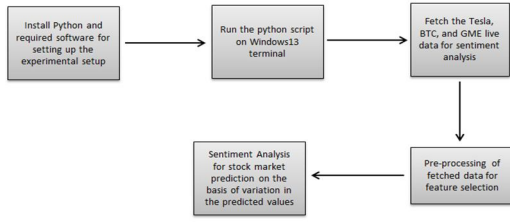


Figure 4: Flowchart for Implementation

```

def scrape_and_process(URLs):
    Article = []
    for URL in URLs:
        r = requests.get(URL)
        soup = BeautifulSoup(r.text, 'html.parser')
        paragraphs = soup.find_all('p') #p is for paragraph
        text = [paragraph.text for paragraph in paragraphs]
        words = ' '.join(text).split(' ')[0:350]
        ARTICLE = ' '.join(words)
        Article.append(ARTICLE)
    return Article
  
```

Figure 5: Code for a Python script that retrieves current server data

### B. Output File

In our study we have collected live URL's of various articles related to stock publication, website news etc in textual form of GME, Tesla and BTC stocks and then sorted the URL's according to our utilization. In next step we have made a list of URL and then we have taken one - one paragraph of article from each URL. Further we have changed this paragraph of article into a line. At last we came across the sentiment of each line by sentiment score through sentiment analyzer. It gave us a final output of our study in term of positive or negative emotion of the stock. Table 1 lists the stock markets that were chosen for this study along with some real time news analysis. The sentiment analysis (positive or negative) performed to anticipate the stock growth rate using real-time data is displayed in Table 1.

TABLE I: SAMPLE SENTIMENT CATEGORIZATION OF REAL-TIME NEWS BY THE MODEL

| Ticker | Summary                                                                                                                     | Label    |
|--------|-----------------------------------------------------------------------------------------------------------------------------|----------|
| GME    | Markets now pricing in a roughly 88% probability of another hike.                                                           | Positive |
| GME    | Fed meeting, existing home sales, Nike earnings.                                                                            | Negative |
| GME    | All images are copyrighted.                                                                                                 | Negative |
| TSLA   | Berenberg downgrades Tesla to Hold, citing valuation.                                                                       | Negative |
| TSLA   | EV-maker is expected to reveal 'generation 3' platform.                                                                     | Positive |
| TSLA   | EV maker says efforts to keep costs low have boosted ability to repay. Tesla bonds now rated Baa3 after S&P in October 2022 | Negative |
| BTC    | Most Asian equity markets gain on relief rally in U.S. equities.                                                            | Positive |
| BTC    | Coin has risen 21% this month, while a choppy S&P 500 and gold have fallen.                                                 | Negative |
| BTC    | We are aware of the issue and are working to resolve it.                                                                    | Positive |

### C. Stock prediction

Sentiment analysis is used to forecast stock growth rate using real-time data. Tesla, BTC, and GME stocks were categorized for this purpose. The stock forecast accuracy of the classified stocks in the proposed system was demonstrated by the sentiment analyzer, which was utilized to test the sentiments.

We used the sentiment scores from the output of the sentiment analyzer tool to calculate the accuracy for each emotion. This method yields two scores: a negative score and a positive score. Specifically, 49.0% of BTC were positive while 48.9% were negative. According to this observation, 49.0% of GME were positive and 54.5% were negative. In particular, 38.5% and 68.5% of Tesla were negative (Figure 6).

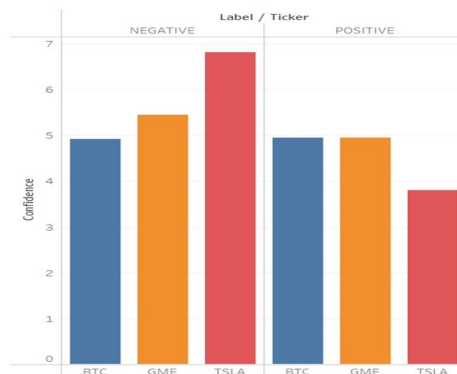


Figure 6: Stock prediction score

## V. CONCLUSIONS

The Indian stock market experiences a number of ups and downs. Making financial investments in the stock market with the intention of purchasing shares requires accurate market forecasting on the part of investors. The primary task for investors is to forecast the state of the stock market, which is dependent on the frequent monitoring and assessment of stock indicator prediction values. By receiving information from the live servers of Tesla, GME, and BTC at different stock news's that may be utilized to forecast the state of the stock market, we have presented sentiment analysis for the stock market in this article. This was done in order to make this achievable. We used the sentiment scores from the output of the sentiment analyzer tool to calculate the accuracy for each emotion which is positive or negative. We decided to use the Python programming language, which has quick execution, to do this. Btc and GME stocks are showing positive emotion, it has higher affirmative sentiments and advised to purchase and will also contribute to the preservation of the share market's economic equilibrium. These Python scripts can be run with more sophisticated functions in the future to complete tasks.

## REFERENCES

- [1] Sun, B., & Ng, V. T. (2014, May). Analyzing sentimental influence of posts on social networks. In Proceedings of the 2014 IEEE 18th international conference on computer supported cooperative work in design (CSCWD) (pp. 546-551). IEEE.
- [2] Liu, B. (2022). Sentiment analysis and opinion mining. Springer Nature.
- [3] Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Intelligent asset allocation via market sentiment views. *IEEE Computational Intelligence Magazine*, 13(4), 25-34.
- [4] Ritesh, B. R., Chethan, R., & Jani, H. S. (2017). Stock movement prediction using machine learning on news articles. *International Journal on Computer Science and Engineering*, 4(3), 153-155.
- [5] Awais, M., Kotecha, K., Akbarzadeh, S., Pandya, S., Ghayvat, H., Chen, W., & Gope, P. (2018). Smart home anti-theft system.
- [6] Awais, M., Ghayvat, H., Krishnan Pandarathodiyil, A., Nabillah Ghani, W. M., Ramanathan, A., Pandya, S., ...& Faye, I. (2020). Healthcare professional in the loop (HPIL): classification of standard and oral cancer-causing anomalous regions of oral cavity using textural analysis technique in autofluorescence imaging. *Sensors*, 20(20), 5780.
- [7] Barot, V., Kapadia, V., & Pandya, S. (2020). QoS enabled IoT based low cost air quality monitoring system with power consumption optimization. *Cybernetics and Information Technologies*, 20(2), 122-140.
- [8] Alostad, H., & Davulcu, H. (2015, December). Directional prediction of stock prices using breaking news on twitter. In 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (Vol. 1, pp. 523-530). IEEE.
- [9] Zubair, S., & Cios, K. J. (2015, January). Extracting news sentiment and establishing its relationship with the S&P 500 index. In 2015 48th Hawaii International Conference on System Sciences (pp. 969-975). IEEE.
- [10] Thanh, H. T., & Meesad, P. (2014). Stock market trend prediction based on text mining of corporate web and time series data. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 18(1), 22-31.
- [11] Shriwas, J., & Farzana, S. (2014). Using text mining and rule based technique for prediction of stock market price. *International Journal of Emerging Technology and Advanced Engineering*, 4(1), 245-250.
- [12] Bing, L., Chan, K. C., & Ou, C. (2014, November). Public sentiment analysis in Twitter data for prediction of a company's stock price movements. In 2014 IEEE 11th International Conference on e-Business Engineering (pp. 232-239). IEEE.
- [13] Cakra, Y. E., & Trisedya, B. D. (2015, October). Stock price prediction using linear regression based on sentiment analysis. In 2015 international conference on advanced computer science and information systems (ICACSIS) (pp. 147-154). IEEE.
- [14] Koukaras, P., Nousi, C., & Tjortjis, C. (2022, May). Stock market prediction using microblogging sentiment analysis and machine learning. In *Telecom* (Vol. 3, No. 2, pp. 358-378). MDPI.
- [15] Idrees, A., Ibrahim, M., & Hegazy, N. (2019). A proposed model for predicting stock market behavior based on detecting fake news. *Empowering Science and Mathematics for Global Competitiveness*, 595-601.
- [16] Desai, M., & Mehta, M. A. (2016, April). Techniques for sentiment analysis of Twitter data: A comprehensive survey. In 2016 international conference on computing, communication and automation (ICCCA) (pp. 149-154). IEEE.