Analyse Stock Price Future Behaviour Using News Sentiments

Raghavendra B [1], M N Priyanka [2], Swathi T [3], Tejaswini K J [4], Yashaswini [5]

[1] Assistant Prof, Department of Computer Science and Engineering, Nagarjuna College of Engineering and Technology, Bangalore, India

[2,3,4,5] B.E. Students, Department of Computer Science and Engineering, Nagarjuna College of Engineering and Technology, Bangalore, India

ABSTRACT-As many people are interested in betting on stock prices and people, the research and prediction of stock values using big data, including artificial intelligence, is now being explored. We performed sentiment analysis in this study by stacking and comparing emotional dictionaries to news items utilising deep learning models. We can determine the optimistic index value of news stories for each day using the emotional dictionary. We can determine the extent and relevance of dynamic analysis in the stock market by examining the correlation value between the positive index value and the stock return value. We compiled a dataset of everyday stock prices of S&P500 businesses during five years and over 265,000 financial news stories on these firms.

Keywords-Sentimental analysis, Sentimental dictionary, Positive index value, Stock return value.

I. INTRODUCTION

Machine learning techniques are being investigated in a variety of sectors nowadays. For stock market participants, accurate forecasting of stock price movement may result in significant profit margins. However, owing to the complexity of stock market data, developing an efficient and successful model for forecasting is quite tricky. Additionally, we may do research using several data sources. Due to the large number of components that contribute to stock price movement, numerous machine learning methods are being investigated, including Bayesian text classifiers and Artificial Neural Networks. Numerous papers evaluate the results of the devices mentioned above.

Additionally, sentiment analysis using Natural Language Processing (NLP) is possible without using different machine learning algorithms. We may use sentiment analysis to develop models for categorizing "tweets" into sound, negative, and neutral attitudes and investigate the thoughts, reviews, and issues raised about the product in blog articles. As a result, sentiment analysis powered by Natural Language Processing (NLP) may be used in various facets of our everyday life. Numerous variables affect stock market pricing. One of these variables is an investor's response to financial news and current events. The availability of news has expanded tremendously in the modern era. Due to the massive volume of news, it is difficult for investors to determine the trend in stock prices. As a result, an automated system that forecasts future stock values benefits investors. Through dynamic analysis, we can demonstrate if the data affected stock prices in this study.

Due to the size of the data sets, they are challenging to evaluate using typical data processing programmes. We can apply machine learning to extract relevant information and convert it to an intelligible structure through Deep Learning algorithms and Sentiment analysis.

As part of our analysis, we develop and simulate a trading system and do market analysis. Researchers have attempted to forecast stock prices using either past stock price information alone or a combination of textual and historical data for years. Several prior types of research employed textual data from Twitter emotions, financial blogs, or news stories. To prevent the spread of false news on social media, we utilise financial news pieces from reputable sources in this study.

We forecasted the current day's closing stock price using historical stock prices plus current financial news. This strategy, we feel, is superior since financial news about the firm has a substantial impact on its stock price. Thus, evaluating a company's financial news rather than its historical stock prices might result in more accurate forecast results.

II. LITERATURE SURVEY

Rubi Gupta et al. [1] conducted sentimental analysis on StockTwits data to understand the impacts of stock price variation. They analyzed only five companies. Using three machine learning methods like Naïve Bayes, SVM, and logistic regression, they got accuracy between 75% to 85%.

Yuguang Huang et al. [2], in training an extensive sample data set, the Naive Bayes algorithm provides accurate result. They investigated the Naive Bayes Classification method using Poisson distribution models and experimental data. They concluded from the experimental findings that this strategy produces excellent classification results in big data sets and produces superior classification results for small data sets.

Gang Li et al. [3] analysed online reviews, blogs, and comments. They then extracted features using the TF-IDF technique. They proposed two techniques to enhance the results: voting and a distance measuring methodology. Then they used K-Means clustering to determine if the data were positive, negative, or neutral.

According to Liza et al. [4], text mining should be divided into three stages: pre-processing, processing, and validation. After finishing the pre-processing phase, they applied weighting schemes and classified them using the Naive Bayes algorithm. Then, during the validation step, they performed tenfold cross-validation to determine the accuracy level.

Ishitha et al. [5] proposed using machine learning technology to forecast the future values of the company's financial stocks. Then, stock values were predicted using regression and LSTM-based machine learning. Following that, they recommended that sentiment analysis by machine learning also improves performance.

Radhu et al. [6] developed a novel method for stock market forecasting. PCASVM was used to eliminate incorrect predictions and determine which characteristics are significant. And found that, when compared to basic SVM algorithms, the created GASVM and PCASVM give significantly improved accuracy and outcomes.

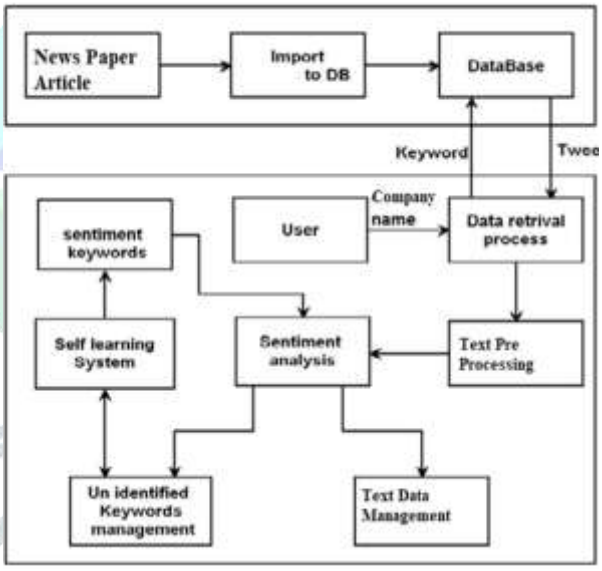
Mohan et al. [7] used three different technologies to forecast stock values: time series models, neural networks, and a mix of neural networks and economic news items. They developed ARIMA, RNN, and Facebook Prophet prediction models. And discovered that the RNN model produced superior outcomes.

Wang et al. [8] shown that using NN (Neural Networks) as a case learning-based technique, news feelings connected to the stock market may be leveraged to enhance the predictor's performance.

Venkata Pagolu et al. [9] used two distinct textual representations, Word2vec and N-gram, to analyse public opinions expressed in tweets. Then, they used sentiment analysis or supervised machine learning methods to the tweets and discovered a link between the company's stock market and tweet sentiment. They demonstrate in this study that a high connection exists between the rise and fall of a company's stock price and the public opinion expressed in tweets.

Sunil Kumar et al. [10] conducted sentiment analysis on Twitter and stock twit's data. Twitter categorizes tweets into four categories: happy, up, down, and rejected. To forecast outcomes, polarity indexes and market data are fed into artificial neural networks.

III. SYSTEM ARCHITECTURE



A system architecture is a model that contains specific information about the system's principles to assist the user in comprehending or familiarising themselves with the topic depicted by the model. This model defines the system's behaviour, structure, and a variety of other aspects. An architectural explanation is a formal description and representation of a system. It is structured in such a manner that it facilitates the system's structure and behaviour reasoning. System architecture may include both the system's components and the created subsystems. The entire system is realised by the combined efforts of subsystems and system components.

As seen in the above-proposed system diagram, the consumer firstly logs in and inputs the firm name of his choosing to receive a stock price forecast.

Initially, data is gathered from news articles, social media, and other sources and then incorporated into databases. The data is acquired from the database using the data retrieval process. Once the data is received or retrieved, it is pre-processed. The term "text pre-processing" refers to the act of removing noise from the raw text in addition to making it predictable and analyzable for our job.

This is a data mining approach that converts collected raw data into useable data in an efficient way. This procedure consists of four basic steps:

1. Data Cleaning

2. Data Integration

3. Data Reduction

4. Data Transformation

Sentiment analysis is used to categorize this pre-processed data into positive, negative, or neutral words or phrases.

Sentimental analysis is a computer technique that identifies and categorizes text based on the thoughts stated in them to determine the writer's attitude, whether positive, negative, or neutral.

Tokenization, data cleansing, and categorization are all steps in this process. Tokenization is separating paragraphs into distinct sets of phrases or phrases into distinct sets of words. Cleaning data is deleting superfluous words, including special characters from raw data, to make it more usable and efficient. Classification is a process of determining if a piece of writing is positive, negative, or neutral. The classification procedure in this suggested system is carried out utilising the naive Bayes classifier method.

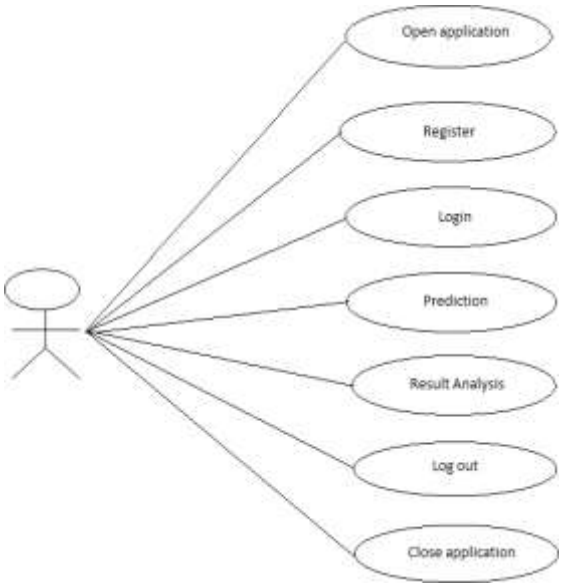


Fig 1: Use case diagram of Stock Price Prediction

We get two sorts of outputs from the dynamic analysis: recognised emotional keywords or unidentified keywords. Sentimental keywords are a collection of words classified as good, harmful, or neutral. These sentiment keywords are managed as text data. It is a method that manages data that is required to run automated tests. On the other side, using a self-learning algorithm, unlabeled data is turned into emotive keywords.

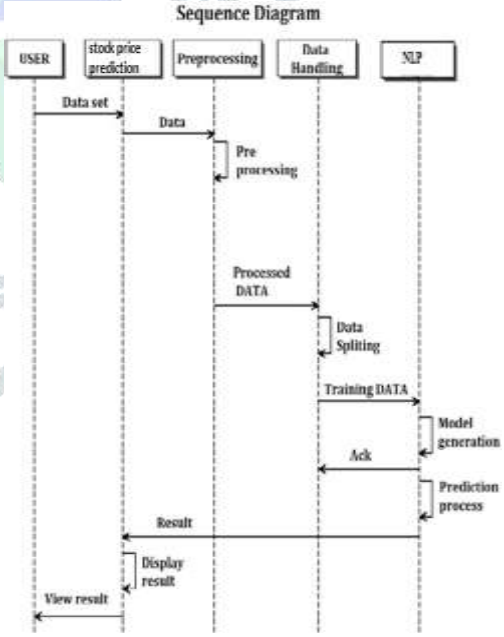


Fig 2: Sequence diagram of Stock Price Prediction

Flask

Flask is a Python web framework. Flask does not need any specific libraries or tools, which is why it is referred to as a micro-framework. It does not include form validation, a database abstraction layer, or any other components that rely on pre-existing third-party libraries to perform everyday

duties. The projected graph derived from the input is incorporated in the application's front-end using Flask.

Python shell and idle

Python is an interpreted language. The term "interpreted language" refers to a language translated step by step into machine-understandable code, which implies that once the first line of the programme is translated and executed, control moves to the following line of the programme.

IDLE is Python's de-facto standard development environment. The name IDLE stands for "Integrated Development Environment." The python shell window provides access to the Python interactive window. Additionally, python source files may be produced and updated using the IDLE's built-in file editor.

To invoke the Python interpreter, enter Python on the climate command line. You'll be presented with a prompt and may begin entering Python instructions. Type ctrl-d to quit the Python interpreter.

PYCHARM

It is a very efficient Python integrated development environment. JetBrains, a Czech firm, created Pycharm. It comes with a visual debugger, code analysis, fully integrated unit testing, and version control system integration (VCs). Additionally, it facilitates web development using Django as data science using Anaconda.

NLTK

The term NLTK stands for "Natural Language Tool Kit." It is both a library and a suite of programmes. This software development kit assists computers in comprehending and analysing human language. Natural language is a field of computer science, most precisely artificial intelligence. NLP is a machine-learning technique that enables robots to comprehend, evaluate, and infer meaning from human languages such as English, Kannada, and French.

IV. NAVIE NAYE'S ALGORITHM

Sentimental analysis is the process used to check whether the given piece of text is negative, positive or neutral. In-text analytics NLP and ML are used together to assign scores of sentiments to categories, topics or entities inside a phrase.

Inexperienced Bayes is a method for machine learning is based upon on Bayes theorem.

Bayes: It is named Bayes because it is founded on the Bayes theorem.

Bayes' theorem: Bayes' theorem, sometimes referred to as Bayes' rule or Bayes' law, is a mathematical method for estimating the probability of a hypothesis given previous knowledge. Conditional probability is used to determine it.

The formula for Bayes theorem is: P(A)=(P(B/A) P(A))/P(B)

where P(A) is a posterior probability, P(B/A) is likelihood probability, P(B) is the marginal probability

Working of Naive Bayes classifier:

The following example helps us understand how the Nave Bayes' Classifier works:

Assuming we have a weather dataset and an objective variable named "Play." As a result, we must use this information to decide whether to play on a particular day, dependent on the weather circumstances. To resolve this impasse, the following procedures must be taken:

1. Create frequency tables from the provided dataset.

2. Create a Likelihood table by calculating the probabilities associated with the provided attributes.

3. Using Bayes' theorem, compute the posterior probability.

V. RESULT

The emotion dictionary is made up of words and their associated positive index (PI), sorted in decreasing order by their frequency. We'll determine which terms are often utilized in news stories. Thus, a large number of terms included in the dictionary. Those words may include words with a lower frequency. Because it was difficult to determine the positive index in such a scenario, the experiment was undertaken by altering the dictionary's word count. The blue line in the accompanying graph represents the line drawn from the actual values, while the black line represents the line drawn from the anticipated values. This is accomplished only via the use of historical values.

Fig 3: Graph of Historical Values

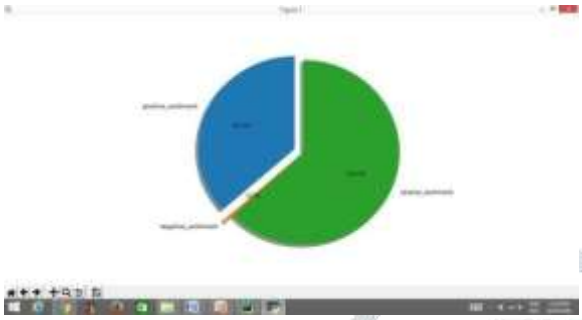


Fig 4: Result Page

The expected outcome is shown to the user in the form of a pie chart.

VI. CONCLUSION

This paper applies two approaches to the Yahoo Finance database: LSTM and regression. There is an improvement in predictive performance, which would be a positive consequence. This experiment illustrates that when machine learning techniques are employed, stock market predictions can be more successful and exact. We have advised in this project that data be acquired from the world's largest financial markets be used. Machine learning algorithms are used to forecast index changes. We used the LSTM algorithm to organise and manage massive data collections. LSTM does not exhibit the fitting problem. The numerical results demonstrate this project's exceptional efficiency. In our research, we hypothesised that personal feeling toward a company's stock might be forecast. This allows us to profit from a stock investment. Due to the unpredictable nature of the stock market and the fact that several factors affect the price of a company, this research is only capable of predicting with a restricted degree of accuracy.

REFERENCES

[1] Gupta, R., & Chen, M. (2020, August). Sentiment Analysis for Stock Price Prediction. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp. 213-218). IEEE.

[2] Zheng, H., & Yi-fang, L. (2017). Market Environment, Stock Price Informativeness and Capital Allocation Efficiency. *Taxation and Economy*, 04.

[3] Li, G., Law, R., Vu, H. Q., Rong, J., & Zhao, X. R. (2015). Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism Management*, *46*, 311-321.

[4] Shoemark, P., Liza, F. F., Nguyen, D., Hale, S., & McGillivray, B. (2019, November). Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 66-76).

[5] Kapoor, I., & Mishra, A. (2018). Automated classification method for early diagnosis of alopecia using machine learning. *Procedia computer science*, *132*, 437-443.

[6] Raghu, S., & Sriraam, N. (2018). Classification of focal and non-focal EEG signals using neighbourhood component analysis and machine learning algorithms. *Expert Systems with Applications*, *113*, 18-32.

[7] Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019, April). Stock price prediction using news sentiment analysis. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 205-208). IEEE.

[8] Wang, D., & Huang, J. (2002). Adaptive neural network control for a class of uncertain nonlinear systems in pure-feedback form. *Automatica*, *38*(8), 1365-1372.

[9] Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016, October). Sentiment analysis of Twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)* (pp. 1345-1350). IEEE.

[10] Kumar, S., Gupta, R., Khanna, N., Chaudhury, S., & Joshi, S. D. (2007). Text extraction and document image segmentation using matched wavelets and MRF model. *IEEE Transactions on Image Processing*, *16*(8), 2117-2128.