International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 09 | Sep 2021 www.irjet.net p-ISSN: 2395-0072

Forecasting Stock Market Trends Using News Headline Analysis

Sukriti Jaitly[1], Mohak Verma[2]

[1,2]*Student, Vellore Institute of Technology, Vellore, India*

Abstract - *This is a study that aims to find a correlation between news headlines and their effect on stock market trends using sentiment analysis, logistic regression, XGBoost, and deep learning. This is a study that aims to find a correlation between headlines of newspapers and their subsequent effect on stock market trends using sentiment analysis, logistic regression, XGBoost, and deep learning. In this study, we suggested a forecasting model for predicting sentiment around stock prices. We map feelings to see if there's a link between news-predicted sentiment and the original stock price, as well as to test the efficient market theory. Finding future stock trends is a difficult endeavor since stock trends are influenced by a variety of factors. Presumably, news items and stock prices are connected. Furthermore, news has the potential to change market patterns. As a result, we set out to investigate this link in-depth and see if stock movements can be forecast using news articles and prior price histories*

Key Words: sentiment analysis, logistic regression, XGBoost, stock market analysis, tokenization, lemming, deep learning.

1. INTRODUCTION

Analyzing and establishing the future estimates and appraisal of stocks is known as stock market prediction, and it is extremely important to the whole financial business. The behavior of human investors in terms of capital pumped into a stock determines the price of a particular stock, investors and research analysts evaluate the market and make strategic decisions to purchase or sell. [13] In recent years, it has drawn the attention of researchers to capture its volatility and predict its trends to help investors and market analysts to evaluate market behavior and organize their investment strategy appropriately.

Several factors influence stock trends, one of which is daily news stories, which may have a significant impact on the fluctuation price of the stock as individuals respond to the given news, Nowadays, big influential business tycoons set market trends by publicly criticizing or supporting, daily news articles and platforms serve the purpose of circulating said information to the public and it can influence trading strategies of the stock market. Therefore, it has become necessary to deeply analyze the information to support the investors to make smart trading decisions before making real investments. The previous study has established a link between news stories and change in flux of stock movements since there is a delay allying with when news is released and published and how the stock market moves to reflect changes in the value of stocks [3].

[7-9] Stock market forecasting provides excellent profit opportunities and is a fundamental catalyst for most researchers in this sector. [15] Most researchers use technical or fundamental analysis to predict the market. We are using unquantifiable data in our work, such as news article headlines, to predict stock market trading, which helps us establish a relationship between the two and find trends that can help traders and variety. We built a sentiment analysis model to classify our unquantifiable data and used logistic regression accuracy to create several predictive models such as Logistic Regression, XGBoost, and others. We've used this to draw a comparison between them and selected the best performer. The feedback from the sentiment analysis model is used as the data for the predictive model.

2. SOCIAL RELEVENCE

One of the most common areas of interest for text mining is stock market analysis. Many researchers have proposed different approaches for predicting the movement of stock market indicators based on text data. Many of these approaches depend on either maximizing the model's predictive accuracy or devising alternative methods for model evaluation. [1] Analysts generally look up information about the companies they are looking to invest in, for long- and short-term trading. Generally, it is noticed that novice traders often follow trends of big investment companies, this makes them risk their capital on a volatile market without any access to informed insights into the market. [4] News sources can provide insight into a company's activities, such as expansion, revenue projection, public sentiment, new products, etc. Depending on the news, traders can determine a bearish/bullish trends and make investment decisions. The existence of news has the ability to influence views or feelings towards specific businesses or business strategies.

This research aims to anticipate future stock trends using non-quantifiable data such as headlines from news articles, as well as news sentiment classification. as headlines from news articles, as well as news sentiment classification. We created a text classification system to classify news stories and anticipate how the news will affect the stock price. This is an attempt to investigate the correlation between stock market patterns and headlines, assuming that news stories have an effect on the stock market.

3. RELATED WORK

[2][6] Stock trend forecasting is a vital and dynamic research area that necessitates exact predictions. Therefore, in recent years noteworthy efforts are put into developing prediction models for the overall stock market. Few of the researchers have shown a strong relationship between the daily news articles and stock prices belonging to a company. In this section, previous research works are elaborated to understand their techniques and feature processing methods.

G. Gidofalvi trained a naive Bayesian text classifier by examining news items that used an approach that included price change and the negative value, which reflects the stock's volatility. Based on the stock's movement vs. its forecasted movement, the articles were labeled in one of three ways. Despite the fact that there was a high link between the news item and stock price behavior during a 20-minute window around the time of the news article's release, the classifier's predictive potential was limited when using this strategy. Fung et al suggested an approach that employed a t-test-based piecewise linear approximation approach. The optimization problem was solved using a Support Vector Machine, with the features being the words in a document weighted using term frequency-inverse document frequency. The strategy that resulted was the most reliable and suitable for predictions within 3-5 days.

4. DATA DESCRIPTION AND EXPLORATORY ANALYSIS

The first dataset is called *Redditnews.csv* – which contains news headlines and dates of 73608 news articles dated from 08/08/2008 to 07/01/2016. Reddit calls itself the "front page of the Internet" and users can upload material anonymously or under their registered username on this social news aggregation, online content rating, and debate platform. Registered users can then vote up or down on entries to determine where they appear on the page. The most popular posts display on the home page or at the top of a category. Reddit contains millions of channels, each known as subreddits, with various topics. Our data set was created with the information crawled from a 'subreddit' called World News Channel. The top 25 headlines for each day in our stated date range were pulled and put into Excel with each headline serving as its own attribute.

The second dataset called *upload_DJIA_table.csv* contains the opening, closing, high, low stock prices dated for the same time period barring the weekend (Stock Market Closure). The Dow Jones Industrial Average (DJIA) is a price-weighted mean of 30 prominent equities listed on the NYSE and NASDAQ. It is a thirty-component index of grouping stocks that is meant to be a gauge or indication of how the overall stock market is performing. Stocks with higher share prices are given greater weight in the index.

4.1 Data Cleaning and Preprocessing

Data cleaning is the most critical and time-consuming piece of any analytics project. Ensuring that the data that is being tested is accurate and clean will lead to better results and reliable processes that can be duplicated/replicated by others that wish to test the validity of the results presented. [7] The process of data cleaning consists of detecting corrupt/ inaccurate records, removing inaccurate or irrelevant parts of data, and coming up with an overall standardization of the data so that algorithms can produce correct results.

Both datasets contain huge amounts of information. At the beginning of our project, we made sure that the data is of the utmost quality. Since the data was pulled from the Reddit website and dumped into a csv format, it led to numerous issues, especially with the news headlines dataset. We encountered trash letters ("\b" as shown in snippets), as well as missing values and misspelled words. Spellcheck was not an option as any word that was classified as misspelled would require a judgment from our group about whether it was intended to be spelled this way or if it was just an error. Relatively, it had a lot of missing, unreadable, and inconclusive variables which had to be taken care of.

4.2 Merging the Datasets

Merging the two datasets was an essential step. The first step was a combination of both files; we wanted the stock price data to concatenate with the Reddit news headlines data. The famous pandas library proved useful. However, the famous pandas.merge() or pandas.concatenate() functions did not work because both datasets had different indexes.

The news dataset had row-wise data with columns: Date and Headline. Each news headline was a separate row entry adding to a total of 73608 news headlines and every unique date having 25 news headline entries which were not consistent throughout the dataset. The stock dataset had stock features for every unique date adding to a total of 1988 unique entries which excluded stock price for Saturdays and Sundays when the market was closed. We tried different python data structures but eventually settled on using lists because we felt that working with them is the fastest.

The approach here was to store all the news headlines in a list. [3] Next, from the news dataset, we stored the index where the date changed marking the start of the next day to a separate list. These two lists were iterated to create a new dataset having unique dates with 25 news headlines along with stock prices. Our final dataset is called CombinedStockNews.csv with 1988 unique dates and each date has Top 25 News Headlines and Stock Market Features.

4.3 Processing Combined Datasets

In order to have a clean combined data set for further analysis. The first step was to eliminate unwanted characters from news headlines.

The second step was to eliminate missing values – replace them with either '' or Nan and addition of features of stock price variations by creating new columns

1. Net Change (Open Stock Price – Closing Stock Price)
2. Net Increase (Highest Stock Price – Open Stock Price)
3. Net Decrease (Open Stock Price – Lowest Stock Price)
4. Target Variable (1 if stock prices increase, 0 if stock prices decrease)

5. Research Questions

- *Is there an impact of positive/negative news headlines on opening/closing stock prices on any given day?*
- *Can we forecast the stock's movement prices by analyzing the text given in news headlines?*

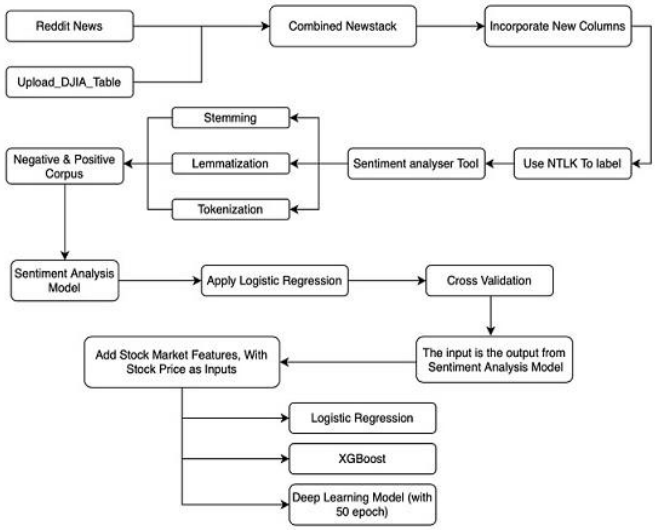

Fig -1: Structure of proposed architecture

6. Methodology

6.1 Labelling

Since the data was unlabeled, the NLTK library was used to label each we utilized NLTK's built-in Vader Sentiment Analyzer, which uses a dictionary of positive and negative terms to rank a text as positive, negative, or neutral. We employed this technique by first constructing a Sentiment Intensity Analyzer (SIA) to categories our headlines, and then calculating sentiment using the polarity scores technique. A negative headline receives a score of 0, while a positive headline receives a score of 1.

6.2 Preprocessing

One of the problems we faced was determining how we could best use the headlines to test against our class attribute. [14] Text mining is still a developing field and there are two main trains of thought about how to best utilize text data. We planned to use sentiment analysis which is the process of determining a sentiment of a sentence based off of a sum score for the sentiment of the words used in that sentence.

Our base assumption was that certain words would have a negative/positive effect on people who invest in the stock market's attitude on how/when they traded thus affecting the DJIA. In conversation and the beginning of the project, we hypothesized that words such as terror, bomb, war, explosion, etc. would have a negative effect on the DJIA. Thus, we needed to include these words in the negative corpus, and positive words into a positive corpus. The last major issue that we were facing in using sentiment analysis and keywords was how to get rid of words that won't affect our analysis/accuracy. This was to be done via tokenization, stemming, and removal of stop words. Stemming or Lemmatization functions would have to be completed on our dataset to remove this issue. Stemming is the process of breaking down words to their root word to avoid having multiple words for what is technically the same word. For example, 'Shoot' would be the root word for 'Shooter', 'Shooting', 'Shootings' etc.

Stop words were terms that were often used but were unlikely to be effective for learning. The process of segmenting text into words, phrases, or sentences is known as tokenization (here we separated words and removed punctuation).

6.3 Model – Sentiment Analysis

A positive and negative corpus was created from news headlines we previously labeled and cleaned. A sentiment analysis model (training – testing) was built to predict the sentiment of a set of new headlines. We used logistic regression, using which we were able to achieve an accuracy of 79.35%.

6.4 Stock prediction Model Building

Multiple predictive models were built to draw a comparison between them and choose the best performing one for future use. The input of this is the output from the sentiment analysis model. Our idea was to use the sentiment analysis model to predict the sentiment of the news headlines for any given day and then use those results along with some stock market features, stock price as the input for the stock prediction model. Hence, we are using multiple models in our analysis. Our best performing models:

Logistic Regression: The accuracy with logistics was around 98.24%.

XGBoost: As XGBoost uses a gradient boosting framework and is very useful in prediction problems involving unstructured data, text in our case. The accuracy with XGBoost was around 98.24%.

Deep Learning: We were able to attain a 97.7% accuracy by employing 50 epochs and a batch size often.

7. CONCLUSIONS

Finding future stock trends is a difficult undertaking because stock trends are influenced by a variety of factors. Presumably, news items and stock prices are connected. Furthermore, news has the potential to change stock patterns. As a result of our thorough investigation, we concluded that stock movements can be forecasted utilizing news items and historical price histories. Automate the sentiment analysis We can estimate an overall news polarity based on keywords in the news stories since news stories reflect opinions about the present market. If the news is good, we may say it has a positive impact on the market, which means the stock price has a better probability of rising. If the news is bad, the stock price may drop in lockstep with the trend

On any given day, there is a hefty correlation between the top news headlines and the changes in the value of stocks bought/sold in the stock market.

Using the sentiment analysis model, we are now capable of predicting the sentiment of a set of news headlines. Using this analysis combined with the stock prediction model, we recommend that the features used, will be beneficial in predicting the fluctuation of the stock market. We would like to expand on this study by including more corporate data and testing the forecast accuracy. We would use Twitter data for comparable research for firms where access to financial news is difficult. We may use similar tactics for algorithmic trading.

REFERENCES

[1] Joshi, Kalyani & N, Bharathi & Rao, Jyothi. (2016). Stock Trend Prediction Using News Sentiment Analysis. International Journal of Computer Science and Information Technology. 8. 67-76. 10.5121/ijcsit.2016.8306.

[2] Lauren, S., & Harlili, S. (2014). Stock trend prediction using simple moving average supported by news classification. 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA), 135-139.

[3] Yu, Wen-Bin et al. "A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting." Int. J. Electron. Bus. Manag. 5 (2007): 211-224.

[4] Tetlock, P. C., Saar-Tsechansky, M. & Macskassy, S, 2008. "More than words: Quantifying Language to Measure Firms' Fundamentals", The Journal of Finance, Volume 63, Number 3, June 2008, pp. 1437-1467

[5] Mittermayr, M.-A., "Forecasting Intraday Stock Price trends with Text Mining techniques", Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004

[6] R.P. Schumaker, Y. Zhang, C. Huang, H. Chen, Evaluating sentiment in financial news articles, Decision Support Systems, 2012, pp. 458–464.

[7] R.P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news: the AZFin text system, ACMTransactions on Information Systems, 2009.

[8] F. Li, The information content of forward-looking statements in corporate filings a naive Bayesian machine learning approach, Journal of Accounting Research, 2010, pp. 49– 102.

[9] W. Antweiler, M.Z. Frank, Is all that talk just noise? The information content of internet stock message boards, Journal of Finance, 2004, pp. 1259–1294.

[10] S.R. Das, M.Y. Chen, Yahoo! for Amazon: sentiment extraction from small talk on the web, Management Science, 2007, pp. 1375–1388.

[11] P.C. Tetlock, M. Saar-Tsechansky, S. Macskassy, More Than Words: Quantifying Language to Measure Firms' Fundamentals, 2008, pp. 1437–1468.

[12] S.S. Groth, J. Muntermann, An intraday market risk management approach based on textual analysis, Decision Support Systems 50 (2011) 680–691.

[13] M. Butler, V. Keselj, Financial forecasting using character N-Gram analysis and readability scores of annual reports, Advances in AI, 2009.

[14] S. Lauren and S. D. Harlili, "Stock trend prediction using simple moving average supported by news classification," 2014 International Conference of Advanced Informatics: Concept, Theory, and Application (ICAICTA), Bandung, 2014, pp. 135-139.

[15] Michael Hagenau, Michael Liebmann, Markus Hedwig, Dirk Neumann, Automated News Reading: Stock Price Prediction Based on Financial News Using Context-Specific Features, HICSS '12 Proceedings, 45th Hawaii International Conference on System Sciences, 2012.