Ilkogretim Online - Elementary Education Online, 2020; Vol 19 (Issue 4): pp. 8497-8502
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2020.04.765237

Twitter Data Sentiment Analysis For Stock Market Prediction Using Machine Learning

Seema Rani, Department of Computer Science & Engineering RDEC, Ghaziabad, India
raniseema@gmail.com

Abstract
Recent outrageous posts on social media have taken the globe by storm and have led to diverse views and views of the general public. Social media plays a significant act for or against a government or a corporation that simply can't decide the movement of market but to grasp the sentiment of twitter data that are posted on social media with good method could be a supreme necessity. It will analyse some twitter postings to grasp human semantic. In any tweet intended posting there are some downgraded keyword. At last, a data-set is ready that consists of unique words collected from twitter posts or comments and so the data-set is trained using Naive Bayes algorithm supported with applied mathematics to spot the sentiment given during a new call and comment . They are going to extract each word of the posting and so it'll be matched by virtue with the data-set words for dilution. Finally, it will be tested to their algorithm using numerous posts from twitter that can deliver the result with good accuracy.

Keywords: Machine Learning; Sentiment Analysis; Stock Market; Naive Bayes classifier, SVM

INTRODUCTION

Systems for predicting the stock market have long been a crucial resource for stock traders.

In general, a variety of factors, including the price of gold, the price of oil, significant events, and last but not least news about stock market businesses, influence the direction in which stocks move. While the majority of parameters taken into account by stock market prediction algorithms are quantitative values, a sizable number of researchers have employed financial news to increase the accuracy of stock direction predictions.

Although the overall accuracy of stock price prediction using historical quantitative data is relatively high [1-3], these approaches are insufficient since they cannot adjust to the price fluctuations brought on by a number of significant events can affect investors' trust since human intuition is lacking. To make up for this deficiency, a number of prediction techniques that take into account both stock market prices and financial news has been improved [4-5]. The findings from many of these investigations, however, do not demonstrate high accuracy. For instance, Schumaker and Chen's [4] suggested method, which relies on noun phrases and proper nouns, only manages to attain accuracy levels of 58% and 58.2%, respectively. Therefore, a superior stock market prediction system needs to be thoroughly researched.

We created a new prediction method called Probabilistic Lexicon Based Stock Market Prediction (PLSP) in order to increase prediction accuracy. The suggested PLSP algorithm forecasts a certain direction for stock price from By calculating the overall probability for each event term from the test data set and applying it to news stories, the likelihood that the stock price will increase or decrease. This report focuses on the closing prices of stocks on the Stock Exchange of Thailand (SET100) and online financial news articles because the study was carried out in Thailand. The experimental data are split into two sets in order to assess the effectiveness of the suggested PLSP algorithm.

Financial news stories from March 2015 to February 2016 make up the first data set utilized to create the suggested probabilistic lexicon. The predictive model was tested, analyzed, and evaluated using news articles from March 2016 to February 2017 from the second data set. The findings show that the suggested model produces superior performance results to alternative models. traditional models: J48, BayesNet, and Support Vector Machine (SVM).

LITERATURE SURVEY

• Predicting Stock Market Prices with Neural Networks, Richard Lawrence In order to forecast stock market prices, neural networks are used in this paper as a survey. Neural networks have the ability to predict market directions more accurately than existing methods because of their capacity to find patterns in nonlinear and chaotic systems. Regression, technical analysis, fundamental analysis, and other popular market analysis methods are explored in relation to the performance of neural networks. Additionally, chaos theory and neural networks are contrasted with the Efficient Market Hypothesis (EMH).

AgarwalApoorva,determine if a tweet is positive or bad, they looked at numerous machine learning algorithms. The author employs a variety of methods, including support vector machines and naive bayes. Support vector machine techniques like the Naive Bayes classifier, which was used to examine sentiment in the tweet data set, would be utilised to forecast market movement.

SVM and SVM derivative have been used in numerous studies to try and increase prediction accuracy. Using various textual representations (bags of words, noun phrases, and named entities) from financial news stories as well as the stock price of the S&P 500, Schumaker and Chen [4] explored stock price prediction algorithms in 2009. The classifier employed was an SVM derivative. They discovered that the model that employed stock price information at the time the news piece was issued and phrases retrieved from the article had the highest directional accuracy, accounting for 57.1% on average. They also thought that was appropriate. Nouns performed better in textual representation than noun phrases, which had a directional accuracy of 58% as opposed to 58.2% for nouns.

A stock price prediction system that took into account correlations between stock market prices and a few words from financial news was put forth by Kaya and Karsligil [5]. The algorithm's key feature was first specified as the word couple (noun and verb) that appeared in the same phrase, and Chi-square was then used to pick features. SVM was employed to categories financial news stories.

From the experiments, they acquired 61% precision and 87% recall. Li et al. [6] suggested another forecasting strategy. Article news and ex-ante pricing were the two data sources employed in this study. As a classifier, SVM was also used in the prediction approach. Considering simply news articles, ex-ante prices, a naive combination of both news articles and ex-ante prices, and four prediction models compared kernel learning (MKL). According to the findings, the MKL model outperformed the other four models in both cross validation and independent testing. The average accuracy of a prediction model combining SVM and MKL in cross validation was 62%, whereas the average accuracy of a model employing SVM and MKL in an independent testing data set was 54%. In [7], the researchers looked at numerous feature types using various feature selection strategies in an effort to increase prediction accuracy. It was suggested to anticipate stock price using context-capturing features.

METHODOLOGY

We used a dataset of historical stock prices for a number of different companies to compare the various machine learning algorithms used for stock prediction. We used a portion of the data to train the algorithms and a different portion to verify their effectiveness. Based on the algorithms' accuracy and mean squared error, we assessed them.

• **DATASET:** Twitter data and Yahoo Finance data were the two types of datasets used.

• **PREDICTION:** A Nave Bayesian algorithm is used for both training and prediction. Data should be processed and normalised before being fed into the Nave Bayesian algorithm. Twitter data is the source of the other dataset utilised for sentiment analysis; this data should also be treated before being used for sentiment analysis.

A. Twitter sentiment analysis

Several avenues for sentiment analysis are provided by research in natural language processing. The first is classification-based.

Gold standard created by humans [8]. To train Nave Bayes or other machine learning algorithms for the analysis of other tweets, all sentiment categories should be supplied in the gold standard [9]. the production of a work by a linguistics group (like Lyashevskaya et al. [11]).

The second method is dictionary-based. We opted to follow Bollen and his colleagues in using a dictionary technique for sentiment analysis because they employed it and had the best results up to this point. Zhang, Fuehres, and Gloor employed this strategy in its most basic form by measuring the number of tweets containing the phrases "hope," "worry," and "fear" [10].

In this investigation, we identify two varieties of lexicon-based methodology. First, we merely count the number of times the words "hope," "worry," and "fear" appear in tweets. Second, we construct more sophisticated dictionaries for each of the eight fundamental emotions and assess the usage of these words. We asked linguistics specialists to develop a gold standard for emotions in tweets so that we could evaluate the effectiveness of emotion recognition. We utilised the conventional metrics of recall, precision, and F-measure to assess the quality of emotion recognition [9].

B. Machine learning algorithms for stock market prediction

Two machine learning techniques that let us categories days based on when events arrive and use the developed model for prediction were employed to test our main hypothesis.

They are Support Vector Machines and Neural Networks.We employ learning techniques on three sets of data to investigate the question: "Does sentiment analysis of tweets provide additional information?" The first collection of data, which we refer to as the basic set (Basic), covered the stock market's features from earlier days. The basic set (Basic&WHF) was supplemented with a normalized amount of tweets that contained the phrases "worry," "hope," and "fear." This generated the second set. The third set was produced by averaging the number of tweets from the following eight types of emotions: "happy," "loving," "calm," "energetic," “fearful”, “angry”, “tired”, “sad”(Basic&8EMO).

Naïve Bayesian Algorithm

One possible computer algorithm addressing differentiation issues is the Naive Bayes algorithm. A supervised learning issue such as a differentiation problem aids in determining the category (sub subpopulation) of a replacement observation from a collection of categories. Some characteristics include the primary observation that is

diagnosed by a special guy, the fact that it is a commonly overused formula and that it has a better depth of understanding to remove data.

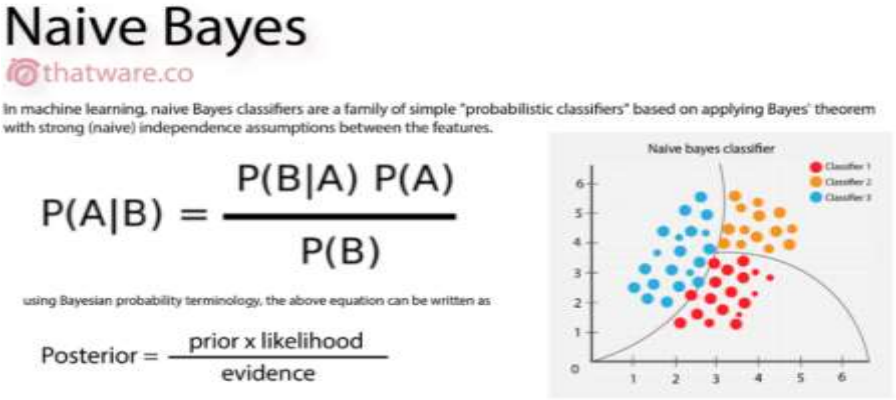

Figure-1

- **Data processing:**

a) To start, we gathered information from Yahoo Finance. A target data frame is formed with simply the close column because the close value is our target value. After that, the data is translated and normalised so that every value falls between 0 and 1.Then, the data is split into two categories: training data (70%), and testing data (30%).

b) Tweety is used to acquire the data from Twitter. The package Tweety is used to connect to the Twitter API. The tweets are cleaned when they are obtained from the API; all links and other special characters are eliminated. After cleaning, they are separated based on their polarity, with positive polarity denoting a positive tweet and negative polarity denoting a

Stock and Data Selection:

For this project Apple, Google, Microsoft's stock values have been used.

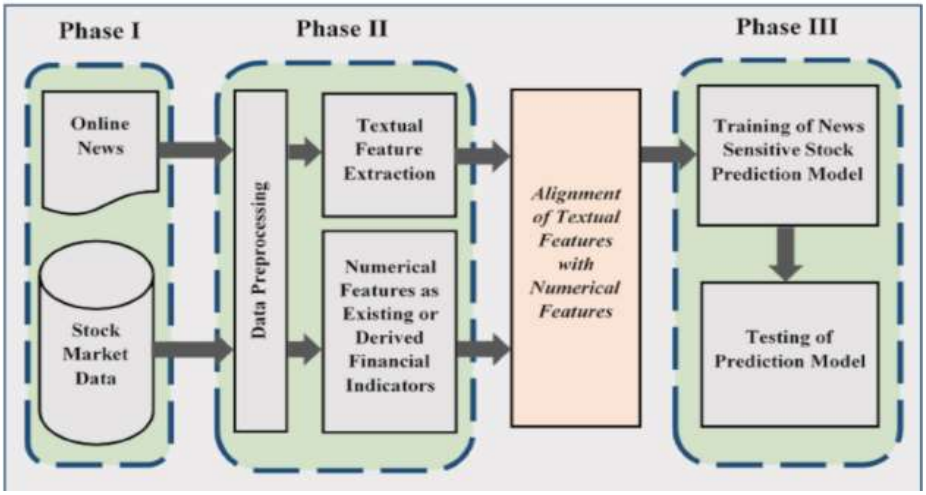

Figure-2

RESULT

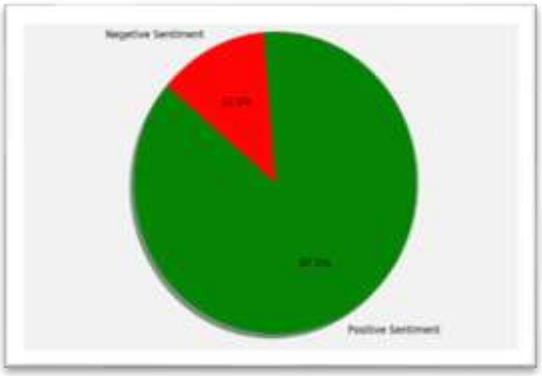

Figure 3: Pie Graph

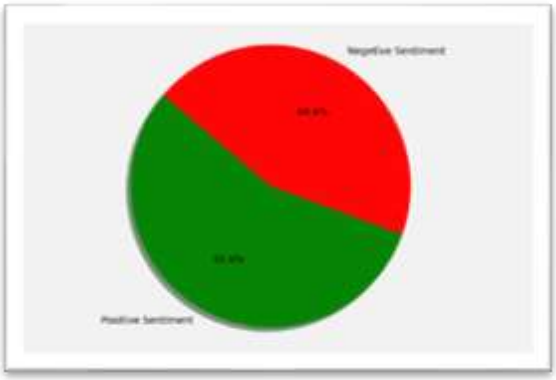

Figure 3: Pie Graph

The GUI will display or forecast the following two sorts of outcomes using our model:
1) The positive feelings in the first pie graph outweigh the negative sentiments, which mean the positive polarity is greater than the negative polarity. As a result, our GUI will advise purchasing the stocks.
2) Because there are more negatively polarised tweets than positively polarised tweets in the second pie chart, our GUI will advise against purchasing this company's stock because it is losing money.

CONCLUSION AND FUTURE WORK

In conclusion, machine learning algorithms can be effective for stock prediction, and each algorithm has its own strengths and weaknesses. The decision tree algorithm is simple and easy to interpret, but it may not capture complex relationships between variables. The artificial neural network algorithm can capture complex relationships between variables, but it may be more difficult to interpret. The support vector machine and random forest algorithms can handle large amounts of data, but they may not be as accurate as the decision tree and artificial neural network algorithms. Future research in this field should focus on developing hybrid algorithms that combine the strengths of different machine learning techniques.The probabilistic lexicon (ThaiFinLex), which is a vocabulary derived from stock market closing prices of the stocks referenced in the news and keywords found in Thai news articles. In contrast to existing lexicons, our proposed ThaiFinLex includes event terms that could affect stock prices and their accompanying probabilities, which forecasts the movement of the stock price. We used split-validation to compare the outcomes of our proposed PLSP model to those of the three other widely used models (SVM, J48, and BayesNet) with 5-fold cross-validation. The outcomes demonstrate that the proposed PLSP model performs better than the

other models taken into consideration in this study. In instance, we were able to get an accuracy of up to 96.64%.Event words with high efficiency typically have an identical effect on all stocks. As an illustration, the phrase "very high profit" has a favorable effect on all equities.

A few terms, however, have varying effects on various equities.

For instance, the phrase "oil prices rose" may favorably impact equities that profit from such an occurrence. However, there is a chance that airline companies will suffer. However, when the EF threshold is 0.7, these terms are removed using the event term analysis. In order to improve prediction performance, we intend to expand the size of the data set and include quantitative analysis models like Moving Average Convergence Divergence (MACD) and Relative Strength Index (RSI).

REFERENCES

1. Mittal, A., & Goel, A. (2012). Stock prediction usingtwittersentiment analysis. StandfordUniversity,CS229(2011http://cs229.stanford.edu/proj2011/StockMarket PredictionUsin gTwitterSentimentAnalysis. pdf),
2. A. Pak,and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," Special Issue of International Journal of Computer replace, France:University ede Paris-Sud,
3. Krouska, A., Troussas, C., & Virvou, M. (2016, July). "The effect of precompleting complicated on Twitter sentiment analysis." In 2016 7th International Conference on call, Intelligence, Systems & replaces (IISA) (pp. 1-5). IEEE.
4. R.P. Schumaker and H. Chen, "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System," ACM Trans. Information Systems, vol. 27, no. 2, 2009, pp. 1–19.
5. M. Y. Kaya and M. E. Karsligil, "Stock price prediction using financial news articles," in Information and Financial Engineering (ICIFE), 2010 2nd IEEE International Conference on. IEEE, 2010, pp. 478–482.
6. X Li, C Wang, J Dong, F Wang, X Deng, and S Zhu, "Improving stock market prediction by integrating both market news and stock prices," in Database and Expert Systems Applications. Springer, 2011, pp. 279–293.
7. M. Hagenau, M. Liebmann, and D. Neumann, "Automated news reading: Stock price prediction based on financial news using contextcapturing features," Decision Support Systems, vol. 55, no. 3, 2013, pp. 685–697.
8. Pang, B., Lee, L., &Vaithyanathan, S. (2002). Thumbs up?:"sentiment classification using machine learning techniques."In Proceedings of the ACL 02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79–86).
9. Paul, M., &Dredze, M. (2011). "You are what you tweet: AnalyzingTwitter for public health."In Fifth International AAAI Conference onWeblogs and Social Media (ICWSM 2011).
10. Mayer, J. D., &Gaschke, Y. N. (1988). "The Experience and Meta-Experience of Mood" 4. Journal of personality and social psychology, 55(1), 102–111.
11. Shah, V. H. (2007). "Machine learning techniques for stock prediction."Courant Institute of Mathematical Science, New York University.