Sentimental Analysis of Tweets using Machine Learning

Akshar Patel[1] Hirenkumar Parmar[2]
[1,2]Devang Patel Institute of Advance Technology and Research (DEPSTAR), India

Abstract— The stock market forecast is a long-term lucrative subject for researchers from various fields. An accurate market prediction is of great importance to investors, but stocks are driven by unpredictable factors, such as microblogs and media, making it impossible to forecast the stocks index based on historical data alone. It is understood that the financial market is knowledge responsive, stock prices represent all the existing knowledge, and price changes can be in response to news or events. For many scholars and economists, the practice of estimating stock markets was a challenging task. Several types of research have been performed in particular to forecast stock market behaviour using algorithms of machine learning. In this paper, we use media platforms and business news data to examine the impact of these data on the accuracy and consistency of stock market forecasts. To accomplish our objective, we used the twitter API and processed it for further analysis. Two machine learning methods were used to evaluate tweet sentiment, namely Naïve Bayes classification and Support vector machines. By comparing each model, we find that the support vector machine provides higher accuracy by cross-validation.

Keywords: Twitter, Yahoo finance, Sentiment analysis, Natural language processing, Naïve Bayes classification, Support vector machines (SVM)

I. INTRODUCTION

The stock market is an integral aspect of the economic development of a country. It is one of the biggest investment prospects for businesses and investors. Only by shares issued to citizens are major capital infusions for businesses around the world. Nation's growth is also strongly tied by our stock market success. Since people's stock market analysis strategies lack expertise and understanding play a very important role in getting new customers on the market and retaining current buyers. To make the right decision to either sell, retain, or purchase other stocks, stock investors must forecast market patterns. Aiming to make a profit, stock investors must purchase stocks whose prices will rise in the foreseeable future and sell stocks whose prices are predicted to decline. If stock investors accurately forecast stock trading patterns, large profits can be achieved. For this reason, stock investors need to forecast upcoming stock market movements.

The stock market is challenging to predict due to its volatile nature, and external influences, such as media platforms and the everyday financial data, have a positive or negative effect on stock prices at once. For specific stock market forecasts, these considerations must be considered. Though investing in the stock market is risky but is amongst the most reliable means of making a meaningful profit when approached in a controlled manner. Thanks to the Internet, shareholders have access to up-to-date news today; the news continuously shapes their sentiments and affects the decision to invest in a specific firm. However, investors cannot adequately judge such an immense volume of social media and financial news. An advanced decision support system is essential for investors because this system automatically assesses market movements using quite large quantities of data. This automated framework can be generated using machine learning algorithms. Finding such a learning machine model that forecasts the stock market correctly would maximize investor profit.

In contemporary days, a significant amount of information is exchanged electronically via numerous social media platforms. Microblogging platforms such as Twitter with a daily login of approximately 100 million people and over 70 billion tweets every day. This engagement can be used as a reflection of the interest of users on a variety of subjects, including the stock market. Although, the behaviour of stock traders and the stock markets are not influenced by social media alone. We can extract essential details that provide a valuable perspective on public sentiment and feelings in some topics. Also, Mathematical techniques have been used to analyze financial data for potential developments in financial markets. Researchers used social media updates or news data in current forecasting frameworks along with stock market data for price prediction purposes. Both data will alter and influence the action of investors, so both data, i.e. social media and financial news, should be considered when implementing a stock market prediction method. Taking into account all sources of prediction the accuracy of the suggested prediction method would improve. This will result in more accurate price market forecasts using social media and financial news. Data sources for the proposed machine learning model would include raw text data in the form of tweets. This raw data cannot be interpreted by machine learning algorithms. The data has to be pre-processed. Natural language processing (NLP) is used for the analytical research approach to interpreting tweets to classify positive, neutral, or negative sentiments depending on the content of texts.

Recently, an important trend in the computing field is to gather, process, and analyze huge amounts of data from social websites to obtain useful information. For this research, we have gathered large quantities of real-time data for twitter, processed tweets and created a feature matrix to implement the algorithm of machine learning. We have also gathered financial data from yahoo finance to predict the stock price. The current research aims to analyze and predict tweet sentiment using Naïve Bayes classifier and Support vector machine models.

II. ANALYSIS AND CLASSIFICATION

A. Data Collection

This subsection discusses the method of data collection, the data collection origins. Tweets related to apple company and apple stock price data are gathered for 2 years from September 17, 2018 to September 17, 2020. Due to its concise nature, twitter is chosen as the source of social media data. Python is used to implement twitter API. As

input parameters, the Python application receives the start date, until to date, output file path, and a search query using cashtags comprising selected stock market ticker symbols followed by $ sign like '$AAPL,'. All tweets from the specified stock markets are imported in raw form in.csv file format between the start date and end date. The downloaded raw tweets file has two attributes one is Tweet text showing the tweet's text, and second is the date, on which the tweet was posted.

B. Preprocessing

The pre-processing of data includes converting raw data into a usable and understandable format that can be transferred for training and testing. It is necessary to pre-process raw data as missing values, incoherent values and much duplication may exist. For a more precise output, we need to solve these problems. The tweets downloaded are in raw form and need to be pre-processed. Each tweet's text includes so many words that do not take into account their sentiments. URLs, descriptions for others and even other marks that have no meaning are used in tweets. In order to get the sentiment of tweet correctly, we must eliminate noise. The first step is to break the text into space, making a list of words per text which is called token. Then by using natural language toolkit we need to remove words from the list which has no sentiment value (for example, is, are, an, etc.).Tweets containing URLs, special symbols(for example,@,$,! etc.) are also removed. On the other hand, the Adjusted column is also dropped for the dataset because the attribute has no part in our stock price forecast model.

C. Feature Extraction and Feature Selection

After pre-processing, the collected data set has several distinctive features. So one extracts the attribute (adjective) from the data set by the feature extraction method. Later this attribute is used in a statement to illustrate the positive or negative polarity which helps evaluate people's views. We retrieve the same features from the tweets to categorise for each classifier. To create a feature set, we analyze each tweet and extracting useful features and then use the unigram method to generate a feature matrix. The Unigram model extracts and distinguishes the attribute. It rejects the previous and subsequent terms in the sentences with the adjective. When the dataset expands, the feature set gets bigger and bigger. It is hard to manage larger datasets after some stage. In this case, unigram is pointless to train the Naïve Bayes and the support vector machine as a feature vector. To avoid such a situation n most significant feature technique is recommended to use. With the chi-squared test, we have identified the best n features from our large set. Chi-squared is one of the standard feature selection methods which selects best n features and removes the remaining features. We classify this feature according to the score after evaluating its feature score and select the top n feature for training.

D. Sentiment Analysis

Due to the abundance of vast amounts of textual data in social media, sentiment analysis has increased in significance in the last decade. For finding users opinions for various domain areas this textual data can be used. There is a considerable significance for sentiment analysis of this vast amount of textual data, data mining and machine learning. Machine learning scholars have therefore been doing experiments on the mining perceptions of users on these platforms. Tweets may be categorised according to their text into various groups. Feature scoring and word count methods have been examined with lexicon-based strategies. Support vector machines (SVM) and Naïve Bayes(NB) have been used for machine learning models. Twitter data was examined and according to the information contained in the text, tweets were classified into three classes- positive, neutral and negative.

E. Training and Testing Models

Training data is used to ensure that patterns in the data are recognised by the computer, cross-validation data is used to ensure that the algorithm used to train the machine is more precise and accurate, and test data is used to see how well the machine can predict new responses based on its training.

We used the AFINN library to plan training sets. It measures positive and negative tweet keywords and the positive, negative or neutral score results in tweet sentiment. We got a positive, negative or neutral keyword map tweet score. If the cumulative score of the tweet is greater than 0, the positive keyword is considered. If the cumulative score of the tweet is less than 0, it is considered negative. If the total score of a tweet is 0, it is called neutral. Tweet date, tweet text, tweet sentiment keyword, and total tweet score are in the created tweet corpus csv file. Using the created tweet corpus csv file, we have prepared datasets.

1) Support Vector Machines (SVM):

Support-vector machines are supervised learning models with related learning algorithms in machine learning that evaluate data used for classification and regression analysis. The SVM Classifier uses a wide classification margin. It divides the tweets with a hyperplane. The data is evaluated by the support vector machine, the decision limits are established and the computation kernels are used in the input space. Two sets of vectors of size m each are the input data. Each data expressed as a vector is then grouped into a class. Then, we observe a distance between the two groups that are far from any document. The distance determines the classifier's margin, eliminating indecisive decisions by maximising the margin. The discriminative function defined below is used by the SVM.

$$g(X)=w^T \phi(X)+b$$

Where,
X= feature vector
w= wights vector
b= bias vector
$\phi()$= non-linear mapping from input space to high dimensional feature space.
Results of SVM:
Accuracy: 83.77 %

CLASS / ERROR	PRECISION	RECALL	F1 SCORE
POSITIVE	0.71	0.55	0.61
NEGATIVE	0.65	1.00	0.81
NEUTRAL	0.93	0.81	0.91

Table 1: Classification report of SVM

2) Naive Bayes Classifier:

In machine learning, naive Bayes classifiers are a family of " probabilistic classifiers "centred on interpreting Bayes'" theorem with clear (naïve) independence assumptions between the features. They are one of the simplest configurations of the Bayesian network models. It has been used in both training and classification processes because of its simplicity. It can learn the trend of evaluating a categorised set of documents. It correlates the contents with the list of words in which the documents are categorised within their correct category. The below equation defines the probability for Naive Bayes

$$P(X \mid y_j)=\prod_{i=1}^{m} P(x_i \mid y_j)$$

Here 'X' is the feature vector specified as $X=\{x_1,x_2,...x_m\}$ and the class label is defined as y_j. There are numerous independent features in our task, such as emojis, sentimental keywords, the number of positive and negative keywords, and the number of positive and negative hashtags are used successfully for classification by the Naive Bayes classifier. The Naïve Bayes is very basic and the expectations of conditional independence in the real world are not practical. We have done it plenty of times it offers greater stock prediction precision.

Results of Naive Bayes Classifier:

Accuracy: 80.12 %

CLASS / ERROR	PRECISION	RECALL	F1 SCORE
POSITIVE	0.45	0.75	0.52
NEGATIVE	0.45	1.00	0.66
NEUTRAL	0.93	0.81	0.87

Table 2: Classification report of Navie Bayes

III. EMPIRICAL RESULT ANALYSIS

Using various assessment criteria, the performance of the chosen classifiers is measured.We used the primary classification metric of accuracy and three different classification metrics within the class, namely, accuracy, recall , and F-score.For examining classifiers, accuracy is a classification metric and it can be represented as.

Accuracy= (Number of correct predictions / Total number of predictions)

If the class distributions in the dataset are not uniform, accuracy will not be a reasonable criterion for the classification success evaluation. Therefore, a confusion matrix is used and its precision, recall and F-score could be used to evaluate classification performance. Precision is the model 's ability to accurately classify samples and it is calculated as follows.

Precision= TP / (TP+FP)

Where TP and FP is true positive rate and false positive rate of algorithm respectively.

Recall demonstrates the potential of the model to identify as many samples as possible and is calculated by the following equation.

Recall= TP / (TP+FN)

Where FN stands for false negative rate of algorithm.

F-score represents both precision and recall and can be shown as follows.

F-Score= 2 x ((Precision × Recall) / (Precision + Recall))

Below, where output classes are compared, a bar chart representation is depicted. We may justify the highest accuracy of the SVM model according to the performance and it is better suited to complete the task.

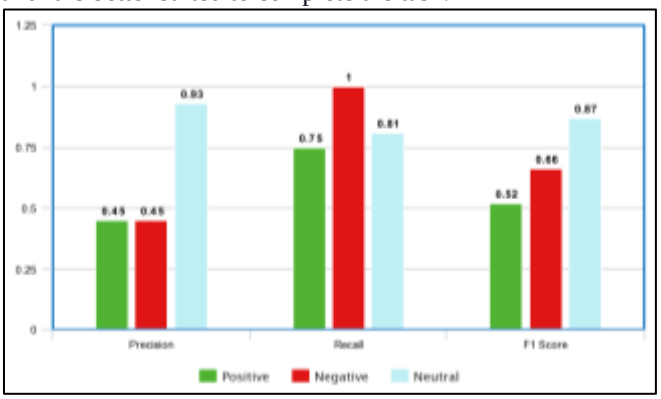

Fig. 1: Comparison of semtiments by SVM

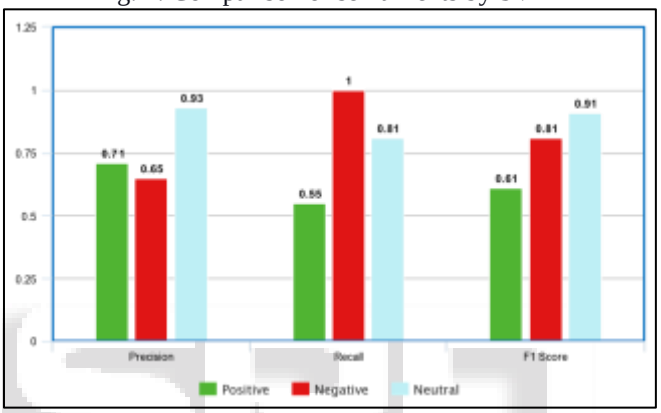

Fig. 2: Comparison of sentiments by Naive Bayes

IV. CONCLUSION

In this paper, we have suggested machine-learning techniques for semantic analysis for classification based on twitter data. To identify sentiments from text, there are distinct Symbolic and Machine Learning techniques. The methods of machine learning are simpler and more effective than symbolic techniques. For Twitter sentiment analysis, these strategies can be used. When dealing with identifying emotional keywords from tweets that have multiple keywords, there are certain problems. Misspellings and slang terms are also hard to manage. An effective feature vector is generated by doing so to deal with these problems. Feature extraction after proper preprocessing in two steps. In the first step, particular features of Twitter are extracted and added to the feature vector. After that, these features are removed from tweets and extraction of features is done again as if it was done on normal text. The feature vector also adds these features. The feature vector's classification accuracy is tested using various classifiers such as Nave Bayes and SVM. For the new feature vector, both of the classifiers have almost similar accuracy.

REFERENCES

[1] B.Agarwal,V.K.Sharma,andN.Mittal,"Sentiment Classification of Review Documents using Phrase Patterns," International Conference on Advances in

Computing, Communications and Informatics (ICACCI), pp. 1577-1580, . 2013.
[2] A.khan,B.Baharudin, "Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms from Blogs," Processed on National Postgraduate Conference (NPC), pp. 1 – 7, 2011.
[3] B.Ren,L.Cheng," Research of Classification System based on Naive Bayes and MetaClass," Second International Conference on Information and Computing Science, ICIC '09, Vol(3), pp. 154 – 156, 2009.
[4] E. Boiy, P. Hens, K. Deschacht, and M.-F. Moens, "Automatic sentiment analysis in on-line text," in Proceedings of the 11th International Conference on Electronic Publishing, pp. 349–360, 2007.
[5] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: A survey," International Journal, vol. 2, no. 6, 2012.
[6] Seng J L, Yang H F. The association between stock price volatility and financial news—A sentiment analysis approach. Kybernetes, 2017, 46(8), pp. 1341–1365.
[7] Afzal H, Mehmood K (2016) Spam filtering of bi-lingual tweets using machine learning. In: IEEE 18th international conference on ICACT, pp 710–714
[8] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment Analysis of Twitter Data," LSM '11 Proceedings of the Workshop on Languages in Social Media, pp. 30–38, Jun. 2011.
[9] Leigh W, Modani N, Purvis R, Roberts T. Stock market trading rule discovery using technical charting heuristics. Expert Systems with Applications, 2002, 23(2), pp. 155–159.
[10] Alostad H, Davulcu H (2015) Directional prediction of stock prices using breaking news on Twitter. In: IEEE/WIC/ACM international conference on WI-IAT 1, pp 523–530
[11] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC, pp. 1320-1326.
[12] Gidofalvi G, Elkan C (2001) Using news articles to predict stock price movements. Department of Computer Science and Engineering, University of California, San Diego
[13] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford 1, 2009.
[14] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.
[15] Brown GW, Cliff MT (2004) Investor sentiment and the near-term stock market. J Empir Financ 11(1):1–27
[16] Yan D, Zhou G, Zhao X, Tian Y, Yang F (2016) Predicting stock using microblog moods. J China Commun 13(8):244–257
[17] Wang F, Zhao Z, Li X, Yu F, Zhang H (2014) Stock volatility prediction using multi-kernel learning based extreme learning machine. In: IEEE joint conference IJCNN, pp 3078–3085
[18] Joshi R, Tekchandani R (2016) Comparative analysis of Twitter data using supervised classifiers. In: IEEE international conference ICICT, 3 pp 1–6
[19] Khatri SK, Srivastava A (2016) Using sentimental analysis in prediction of stock market investment. In: IEEE 5th international conference ICRITO, pp 566–569
[20] Khan W, Malik U, Ghazanfar MA, Azam MA, Alyoubi KH, Alfakeeh AS (2019) Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis.
[21] Bhardwaj A, Narayan Y, Vanraj, Pawan, Maitreyee D. Sentiment analysis for Indian stock market prediction using Sensex and nifty. Procedia Computer Science, 2015, 70, pp. 85–91.
[22] R. Feldman," Techniques and Applications for Sentiment Analysis," Communications of the ACM, Vol. 56 No. 4, pp. 82-89, 2013.
[23] Chen W, Zhang Y, Yeo CK, Lau CT, Lee BS (2017b) Stock market prediction using neural network through news on online social networks. In: IEEE international ISC2, pp 1–6

