Sentimental Analysis of News Headlines for Stock Market

Gaurav Jariwala
Computer Department
Sarvajanik College of Engineering & Technology
Surat, India
gjariwala9@gmail.com

Harshit Agarwal
Computer Department
Sarvajanik College of Engineering & Technology
Surat, India
9arshit@gmail.com

Vraj Jadhav
Computer Department
Sarvajanik College of Engineering & Technology
Surat, India
vrajjadhav0073@gmail.com

***Abstract*—Predicting the stock market is like one of the most common things one does. Predicting exact value is like a futile attempt but one thing is sure we can know the trend or so called the direction in which the price of the stock will move. The stock market is driven by a lot of factors and the majority of them are in the form of news articles. The impact of news on stocks is quite significant. The direction of stock price will be dictated by the sentiments of the market participants. The sentimental analysis is no new topic people have been doing it for quite a long time by implementing several models. In our research paper, we are comparing the results from different models under the same circumstances and concluding which one of the compared models is better based on their accuracy. The methods that we have used for comparison are K-Mean clustering, Naïve Bayes, and Support Vector Machine. In our experimental study for sentimental analysis for news headlines, we found that Support Vector Machine and Naïve Bayes have better accuracy than K-Means clustering.**

Keywords—K-Means Clustering, Moving Average, Naïve Bayes, Support Vector Machine, Sentimental Analysis, Text Classification.

I. Introduction

The stock market nowadays has become an obsession among the millennials. It is a place where anyone can join without much of an effort and the main reason is monetary gain. The two questions that should arise in mind before participating in which major factors affect the change in price and the second question should be at what price one should buy and sell the stock. The major factor that is responsible for the change in price is the sentiment of the market participants and the sentiments changes according to the news, hence by analyzing it, we can know the trend of the market. The ideal scenario is to buy a stock at a low price and sell stock at a higher price for profit. Although it sounds simple, one has to be sure of the position of the current price which can be evaluated by technical analysis of that stock. Technical analysis has been given a lot of importance but we can achieve more convincing results by combining technical and sentimental analysis.

For instance, news for a particular stock states that its quarter is going to be exponentially good due to which the effect of this news on the stock price will be significant in a positive way in the upcoming days and the whole situation can be vice versa. Due to a large amount of data, manually evaluating is tedious and parallel advancement in machine learning can help us achieve this goal easily. There are many machine learning algorithms for performing sentimental analysis out of which we are considering Support Vector Machine, Naïve Bayes and K-mean clustering.

The remaining paper is organized as follows: Section II, summarizes the related work done in this domain; Section III, describes our approach and implementation; Section IV describes the results obtained from the different systems; Section V summarizes the conclusion and future directions of the concept.

II. Related Work

Győző Gidófalvi [1] hereby presents us with an approach to use a news article for anticipating the price trend rather than predicting a precise value. he has used naïve Bayesian text classifier to make classes (up, down, unchanged) then they compared the news article for that specific stock with the price movement of that stock before and after 20 minutes of news article been published. The predictive power of the algorithm was not good but they found a strong link between the news article and the change in the stock price in the time interval of 20 minutes before and 20 minutes after the news article was published.

Pegah Falinouess [2] has used a different approach to label the news from the article and has compared it with the random news labelling. The algorithm they have used is Support Vector Machine (SVM). When the results were compared the accuracy of her model was 83 percent whereas the accuracy of random news labelling was 51 percent.

Yauheniya Shynkevich, T.M. McGinnity, Sonya Coleman and Ammar Belatreche [3] used a method of dividing the news from news article into five categories these categories were based on news relevant to the specific stock, news related to its sub-industry, news relevant to its industry, news relevant to a group of industry and news relevant to its sector. Different types of kernels were used to learn from these subsets. from various

methods used Math kernel library (MKL) proved to be an appropriate one with good results.

Adam Atkins, Mahesan Niranjan and Anrico Gerding [4] explained that the change in volatility was better predicted than the close price of the asset as per there model. Their main aim was to show the relationship between news driven information and implied volatility of the underlying stocks. The implied volatility were derived from the option pricing formula.

III. METHODS

We are using sentiment analysis methods on news headlines for the stock whose next day's movement is to be predicted. The news headlines are extracted from the web using web scrapping from moneycontrol.com using beautifulsoup [5] and selenium [6]. The data is pre-processed and cleaned. After that, each headline is classified as positive, negative, and neutral. Scikit-learn library [7] is used for data analysis and Natural Language Toolkit (NLTK) library [8] is utilized for language processing.

The data consist of the news headlines of the stock and labels to predict if the value of the stock will increase or decrease on the next day. We are a using moving average to calculate the label. Common time periods for moving average are 5 days, 10 days, 15 days, 20 days, 50 days, 100 days and, 200 days for stocks. For our calculation of labels, we have considered the moving average for 5 days, 10 days, and 15 days.

The label for the current day is calculated by comparing these three average values. If the average of 5 days is greater than 10 days, and 10 days is greater than 15 days, then it means the stock value will go up on the day which is indicated by '1'. If that's not true then the label for the current day is kept '0' which indicates that the stock value on the next day will either goes down or will remain the same.

$$L_c = \begin{cases} 1 \text{ if } A_5 > A_{10} > A_{15} \\ 0 \text{ else;} \end{cases} \quad (1)$$

Where: L_c is label is of current headline

:A_5 is average of next 5 days from current day

:A_{10} is average of next 10 days from current day

:A_{15} is average of next 15 days from current day

A. K-means Clustering

K-means clustering is the most basic unsupervised method for data clustering. The process iteratively selects random centroids and based on the Euclidean distance between each data point and the centroid, the data point is assigned to a cluster with the least distance.

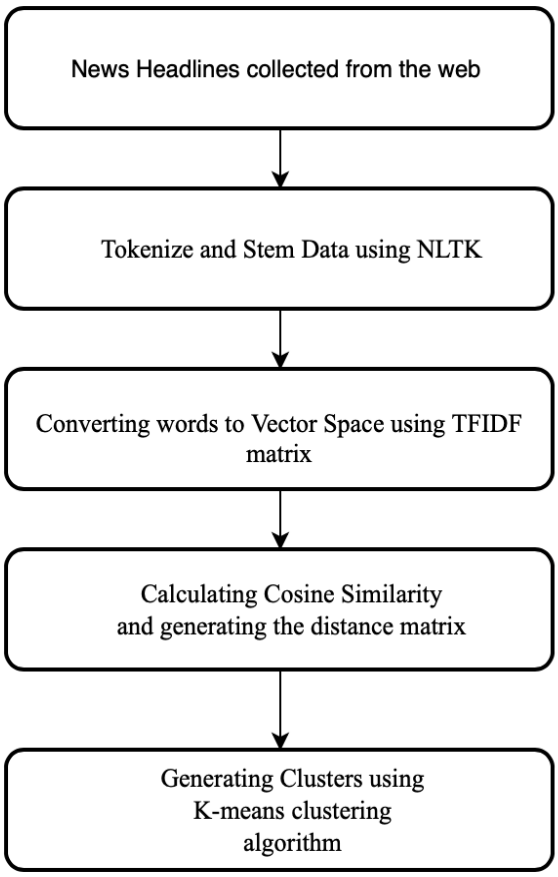

Fig. 1. Flowchart for classification of news headlines using K-means clustering

Word Tokenization helps in splitting the sample sentences into words which are then transformed into their basic form and help in removing the stop words. Cosine Similarity [9] helps us in understanding the similarity of headlines. Tfidfvectorizer helps in computing the word counts, idf and tf-idf values all at once. Finally, clusters are generated and headlines are assigned to a different cluster based on the calculations and data passed.

B. Naïve Bayes

Naïve Bayes (NB) classifier is based on Bayes theorem [10]. It is a supervised learning algorithm. It uses probabilistic learning method for text classification [11]. They require a small amount of dataset to estimate necessary parameters [12]. However, it assumes that features that are independent of each other in a given class make this system suitable only for a specific dataset that follows this rule to some extent [11].

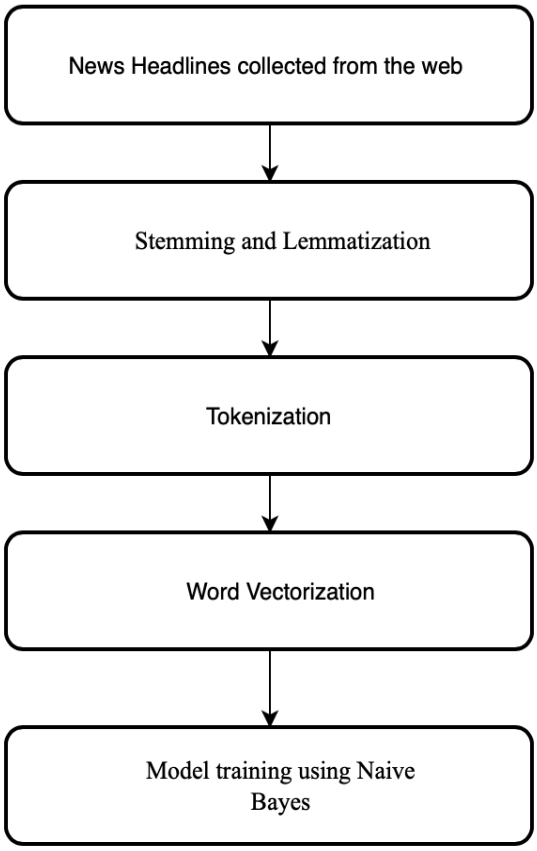

Fig. 2. Flowchart for classification of news headlines using Naïve Bayes

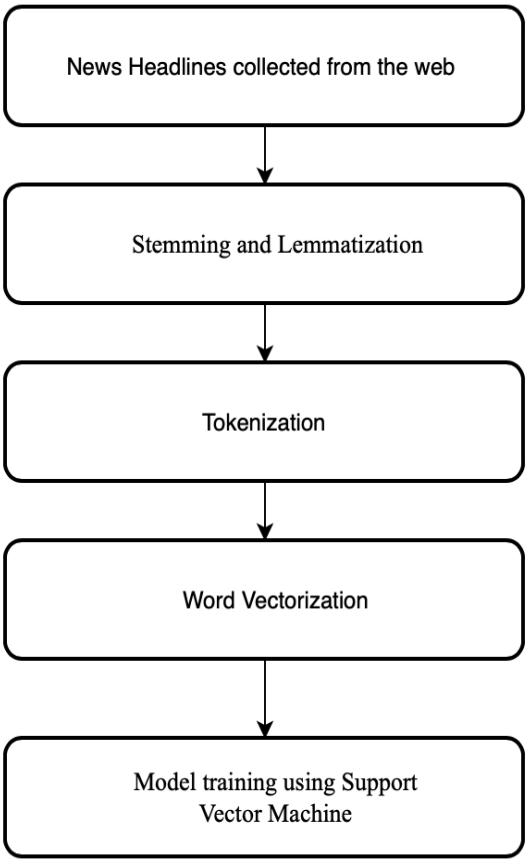

Fig. 3. Flowchart for classification of news headlines using Support Vector Machine

Stemming is a process of converting the words to their base form. Lemmatization groups word with similar meaning in a single group for analysis. These steps help in reducing the redundancy by removing multiple variants of the same word. Tokenization is the process of splitting sentences into words and vectorization is the process of counting the occurrence of the words in the dataset. The pre-processed dataset along with the metadata is passed to the classifier for training and prediction of labels.

C. Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm. SVM creates a hyperplane in N dimensions, which is used to classify data points. The hyperplane is selected on the basis that the distance between the nearest point of each class and hyperplane is maximum. Since there is virtually no limit to the dimensions of the hyperplane, SVM becomes very powerful in extracting features that may not be visible or are too complex to compute for other classification systems.

In our experiment, the data preprocessing steps for SVM are the same as Naïve Bayes system. SVM produces highly accurate results with very less computational power.

IV. Results

In this experimental study, we have trained Support Vector Machine, Naïve Bayes, and K-mean clustering on the news headline of Mahindra and Mahindra Ltd. which comes under the automotive sector, and Kotak Mahindra Bank Ltd which comes under banking sector.

The dataset used to train these models were taken from moneycontrol.com from the period of 2009 to 2019 for both the companies. Fig. 4. shows the sample of the training dataset.

Dates	Headlines	Labels
14-Oct-09	kotak mahindra have target of rs 825: m thacker	1
16-Oct-09	buy kotak mahindra bank, say sukhani	1
27-Oct-09	kotak mah bank q2 con pat up at rs 300 cr	0
10-Nov-09	kotak mahindra bank say credit offtake mute so far	0
18-Nov-09	buy kotak mahindra bank; target rs 870: india infoline	0
23-Nov-09	hold kotak mahindra bank: anu jain	0
08-Jan-10	kotak mahindra bank q3 pat see at rs 98 cr: prabhudas	0
21-Jan-10	kotak mahindra bank q3 net profit 153.09% at rs 331.3 cr	1
21-Jan-10	kotak mah bank q3 net profit see up 120% at rs 287.9 cr	1
25-Jan-10	short kotak mahindra on bounce: anu jain	1
22-Feb-10	hdfc bank have target of rs 1850-1900: mohindar	0
24-Feb-10	keep rs 724 stoploss in kotak mahindra bank: thacker	0
09-Mar-10	sell kotak mahindra bank: gorashekhar	1

Fig. 4. Sample dataset to train the classifiers

TABLE I. CLASSIFICATION ACCURACY USING THE PROPOSED SUPERVISED METHODS

Companies	**Total Headlines Tested**	**Correctly Classified**		**Accuracy**	
		SVM	***NB***	***SVM***	***NB***
Mahindra and Mahindra Ltd.	536	355	356	66.23%	66.42%
Kotak Mahindra Bank Ltd.	251	207	208	82.47%	82.87%

TABLE II. CLASSIFICATION ACCURACY USING THE PROPOSED UNSUPERVISED METHOD (K-MEANS CLUSTERING)

	Mahindra and Mahindra Ltd.	**Kotak Mahindra Bank Ltd.**
Total Headlines Tested	2679	1511
Correctly Classified	1734	876
Accuracy	64.73%	57.97%

Table 1 shows the classification accuracy using the proposed supervised methods. In the proposed supervised methods, total headlines tested were 536 and 251 for Mahindra and Mahindra Ltd and Kotak Mahindra Bank Ltd. respectively. Both SVM and NB have almost similar accuracies where NB had a bit higher accuracy than SVM. For Mahindra and Mahindra Ltd., SVM gave an accuracy of 66.23% while NB gave an accuracy of 66.42%, and for Kotak Mahindra Bank Ltd., the accuracies achieved were 82.47% and 82.87% by SVM and NB respectively.

We performed K-means clustering on the news headlines dataset and the result we achieved is shown in table 2. We tested a total of 2679 and 1511 news headlines for Mahindra and Mahindra Ltd. and Kotak Mahindra Bank Ltd. respectively. The accuracies we received were 64.73% for Mahindra and Mahindra Ltd. and 57.9% for Kotak Mahindra Bank Ltd.

V. Conclusion and Future Work

Unlike the conventional stock market prediction system our novel approach combines the sentiment of market participants through the news feeds and moving average of the stock price. Support Vector Machine and Naïve Bayes are sophisticated enough to analysis news headlines for the stock market. Due to the complexity of the dataset K-means clustering for news classification is not recommended. In several occurrences, news sentiment might be positive but its detailed analysis might contradict its effect on stock's price and since K-means is an unsupervised algorithm, it does not have any means to learn these intricate relationships. Thus supervised algorithms have a better accuracy rate. Naïve Bayes is giving slightly better results compared to SVM as Naïve Bayes considers features independent of each other in a given class.

For future work, global trends and events can be used to further improve the prediction of the stock's direction for a longer timeline. Also by analysis of the entire news article instead of news headlines may give better results.

References

[1] G. GidoÃÅfalvi, "Using News Articles to Predict Stock Price Movements," San Diego, CA: University of California, 2001.

[2] P. Falinouss, "Stock trend prediction using news articles : a text mining approach," Dissertation, 2007, pp. 83-84.

[3] Y. Shynkevich, T. M. McGinnity, S. Coleman and A. Belatreche, "Predicting Stock Price Movements Based on Different Categories of News Articles," 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, 2015, pp. 703-710.

[4] A. Atkins, M. Niranjan and E. Gerding, “Financial news predicts stock market volatility better than close price,” The Journal of Finance and Data Science, 2nd ed., vol. 4, 2018, pp. 120-137.

[5] L Richardson, “Beautiful soup documentation,” April, 2007.

[6] S. Salunke, “Selenium Webdriver in Python: Learn with Examples,” CreateSpace Independent Publishing Platform, North Charleston, SC, USA, 1st. ed., 2014.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos and D. Cournapeau, ”Scikit-learn: Machine learning in Python,” Journal of machine learning research, pp.2825–2830, 2011.

[8] S. Bird, E. Loper and E. Klein, “Natural Language Processing with Python,” O’Reilly Media Inc., 2009.

[9] H. Liang, “Coevolution of political discussion and common ground in web discussion forum,” Social Science Computer Review, 2nd ed., vol. 32, 2014, pp. 155-169.

[10] H. Agarwal and G Jariwala, “Analysis of Process Scheduling Using Neural Network in Operating System,” in Inventive Communication and Computational Technologies. Lecture Notes in Networks and Systems, vol 89, G. Ranganathan, J. Chen, Á. Rocha. Eds, Springer, Singapore, 2020, pp 1003-1014.

[11] P. Kaviani and S. Dhotre, “Short Survey on Naive Bayes Algorithm,” International Journal of Advance Research in Computer Science and Management, vol. 4, 2017.

[12] H. Zhang, “The Optimality of Naive Bayes”, New Brunswick, Canada: University of New Brunswick, 2004.