Sentiment analysis and prediction of Indian stock market amid Covid-19 pandemic

Sentiment analysis and prediction of Indian stock market amid Covid-19 pandemic

Chetan Gondaliya*[1], Ajay Patel[2], Tirthank Shah[3]

[1]*AMPICS, Ganpat University, Mehsana, Gujarat*

[2]*AMPICS, Ganpat University, Mehsana, Gujarat*

[3]*Institute of Management, Nirma University, Ahmedabad, Gujarat*

[1]cpg01@ganpatuniversity.ac.in,[2]ajaykumar.patel@ganpatuniversity.ac.in,[3]tirthanks@gmail.com

Abstract. Outbreak and spread of the Covid-19 pandemic have touched to the core of our sentiments. Indian stock market has seen a roller coaster ride so far this year amid the Covid-19 pandemic. Sentiments have turned out to be a significant influence on the movement of the Indian stock market and pandemic has only added more steam. This study with the limelight on the Covid-19 pandemic is an endeavour to investigate the classification accuracy of selected ML algorithms under natural language processing for sentiment analysis and prediction for the Indian stock market. The study proposes the framework for sentiment analysis and prediction for the Indian stock market where six ML algorithms are put to test. Consequently, the study highlights the superior algorithms based on accuracy results. These superior algorithms can be potent input to build robust prediction models as a logical next step.

Keywords: ML algorithms, Natural language processing, Sentiment analysis, Sentiment prediction, Indian stock market

1. Introduction

Voicing out opinions and expressing reactions have been central to human development. The advent of social media sites such as Twitter, Myspace, Facebook, YouTube etc. and networking sites such as LinkedIn, Xing, Meetup etc. have provided an instant digital platform to express the opinions as well as reaction to any news or announcements in textual format. Such enormous textual data has turned out to be a gold mine to researchers. The Covid-19 pandemic has been centered on generating a huge number of text messages as an expression of various sentiments. These text messages are invariably the writer's expression of opinion or reaction towards pandemic, product, service, people, news, events and other such instances. Analyzing such texts through appropriate techniques to unearth the sentiments of the writer behind it, can lay a strong foundation for predictive models.

'Sentiment analysis' as a term was first used by [1], which has gained significant traction in the research world. Sentiment analysis is widely applied to diverse fields such as movie reviews [2], buying behavior in financial products [3], towards improvement in products and services [4], improving services of government agencies by analyzing customers' review [5]. Stock market

movements in general and price movements of the stocks, in particular, have been an immense source of opinion and reaction generator from several viewers, traders and investors. It is understood that in short-run stock market movements are majorly influenced by sentiments of participants. Many studies have focused on sentiment analysis and prediction based on models in context with stock market movement. Investors' sentiment on the company's earnings announcements has an impact on the price movement of the US stocks [6]. Firm specific sentiments and general market sentiments have been studied to understand and to predict the influence of sentiments on the pricing of selected sample US companies [7]. Analysis of sentiment signals from experts based on Twitter feeds has shown better predictive power for stock market price movements [8].

However, it is to be noted that building a prediction model based on particular sentiment analysis without identifying the optimal algorithm technique to perform such sentiment analysis may lead to misleading inferences. Hence, this paper is an attempt to conduct and compare sentiment analysis and prediction of sentiment on Indian stock market news by applying six algorithms techniques namely; Decision Tree method, Random Forest method, Logistic Regression method, Naïve Bayes method, Support Vector Machine method and KNN method with the aim to identify the most suitable technique for sentiment prediction based on accuracy with special focus on the time period which covers the outbreak and spread of Covid-19 pandemic. This study contributes towards the identification of optimal algorithm techniques to build and test a superior predictive model as the next logical step of research. The sources for text messages to undertake sentiment analysis are identified as forums discussion, financial news, stock market tweets and RSS feed related to the Indian stock market. Different approaches have been used by researchers for sentiment analysis. Machine learning techniques are used for sentiment analysis and prediction based on Twitter feed [9]. The vector space model approach is applied to measure sentiment orientation [10]. While the unsupervised approach is used to conduct sentiment analysis on financial news [11].

This study is divided into sub-sections where the second section highlights the literature review, the third section represents data, methodology and proposed framework, findings and discussion which is heart of any empirical testing is elaborated in the fourth section and concluding remarks and future scope of research is deliberated in the fifth section.

2. Literature review

Sentiment analysis has increasingly received significant research attention. Opinion mining has emerged as a key tool to comprehend the sentiments of the target audience in order to build superior predictive models. An elaborative review of the evolution of sentiment analysis was conducted by [12] with special context to research topics and highly cited papers. Research work in the domain of computer-based sentiment analysis has seen an upsurge since 2004 and application of sentiment analysis in divergent fields like cyberbullying, elections, stock markets, medicine, disasters etc. has been witnessed [12]. As the prediction of stock market movement has always lured many researchers as well as practitioners, it has attracted huge attention from researchers.

[13] has conducted sentiment analysis by using python script language on Indian stock market news concerning predicting the movement of Sensex and Nifty as the stock market index. Like-wise supervised sentiment analysis was used as a tool to predict the buying behavior of participants in the Indian Futures market which is a type of financial derivative [3]. How social media sentiment influences the stock price movement with special reference to Apple Inc. was analyzed and accuracy was estimated [14]. The part-of-speech graphical model was used under model-based opinion mining to test the prediction power in the Iranian stock market based on the text of the opinion of users [15]. A study by [16] suggested the utility of data from microblogging sites for predicting behavior and movement of the stock market where variables such as trading volume, volatility and returns of stock were taken into consideration.

Deep learning approaches or techniques have been adopted by studies such as [2], [4], [16]. The study [2] reported exploration and comparison of various methods under deep learning. As per [4], the deep learning model was used to construct a classifier for detection of sentiment which has reported a higher range of precision in the domain of satisfaction of customer and opportunities identification for products and services improvement. Five different methods of regression; Ensemble Averaging method, Neural Network, Random Forest, Multiple Regression and Support Vector Machine method is used to perform sentiment analysis for assessing the prediction power in the context of the S&P 500 index a stock market index of US [16].To classify sentiment, the machine learning technique of Naive Bayes classification was applied and the study has found that there was an influence of sentiments represented on various platforms on the price of Apple Inc.[14]. Enhanced learning-based method (enhanced NN system) was used by [17] which has reported that the performance of this method can be increased by having selected proper window size. Feature extraction technique based on N-gram was applied to categorize tweets in regards to feature vector while tweet class prediction was conducted based on SVM and Naïve Bayes classifiers for prediction analysis in context with the influence of stock market news on the future movement of stock prices by [18]. A study by [19] has suggested that the SVM model with the segment index has shown accuracy of prediction on the higher side as compared to the SVM model without a combined segment index in regards to using social media text to predict the movement of the stock market.

3. Data set and methodology

3.1. Data

As the data sources for the collection of text messages have become abundant, it is crucial to identify appropriate text sources or platforms from which text messages are selected for sentiment analysis. Four key sources namely; RSS Feed, Forum discussion, Twitter and News portals were selected for extracting opinions and reactions to stock market movement in text format. The data has been collected for the very recent time period from 1st-January-2020 till 24th-August-2020. The Figure-1 describes the data sources which is given below;

Table 1. Data collection sources.

RSS Feed	Forum discussion	Twitter	News portals
Money Control	Money control	Hashtag	Marketmojo
Economics times	Yahoo conversations		Outlook India
United news of India	Traderji		Money Control
Business standard	Valuepickr		Economic times
			United news of India
			Millenniumpost
			Htsyndication
			The Himalayan
			Business standard
			Free press
			Equity bulls
			Business line
			The statesman
			Investing
			Stock adda
			Bse2nse
			Reuters

Labeling of textual data has been done under three categories; Positive, Negative and Neutral. Word association for positive and negative words is done as per Table-2 which is represented as below.

.

Table 2. Word association for labelling under positive and negative.

Sr. No.	Positive phrases	Negative phrases
1	Bullish	Bearish
2	High	Low
3	Outperform	Underperform
4	Surge	Decline

3.2. Methodology

Natural language processing (NLP) is considered to be a powerful tool for understanding and interpreting human languages like speech and text. Natural language processing is useful in text analysis and text mining. Frequently, used three techniques for natural language processing are; 1) Bag of words 2) N-grams 3) TF-IDF (Term Frequency — Inverse Data Frequency)

3.2.1 Bag of words method

Under natural language processing, the bag of words model is used for the representation of text data in a simplified way. Bag of words method works as the name suggests by creating a bag that is represented as a collection of words for particular text such as a document or a sentence. This model is hugely used for document classification.

3.2.2 N-grams method

The N-grams method is quite useful in natural language processing. N-grams model generates document term matrix in which counts are there a part of cell but not as single terms represented in columns but as a representation of adjacent words' combination of particular length 'N' in specific text. The example of how N-grams model works is as under;

Example: "NLP is an interesting topic"

Table 3. N-grams.

n Name	Tokens
2 bigram	["nlp is", "is an", "an interesting", "interesting topic"]
3 trigram	["nlp is an", "is an interesting", "an interesting topic"]
4 four-gram	["nlp is an", "is an interesting", "an interesting topic"]

Formula:

The underlying formula for N-grams model is as under;

If we calculate the probability of 'w1' word coming right after 'w2' word, then the formulation of equation is represented as follows;

count (w2 w1) / count(w2)

it means in a given sequence the number of times word occurs, divided by number of times the word occurs before the expected word in the corpus.

3.2.3 TF-IDF model

TF-IDF model stands for 'Term Frequency — Inverse Data Frequency'.

- Term frequency (TF): Term frequency provides a frequency of the word in each given document in a corpus or collection. Term frequency is represented in ratio format where the number of times the word appears in a given document is numerator while the total sum of words in the document is the denominator. As the occurrences of the word increases, the ratio tends to go up. Every document can have the term frequency of it.

- Inverse Data Frequency (IDF): Inverse Data Frequency is used to measure the rare words' weight in all the documents which are there in the corpus or collection. High Inverse Data Frequency (IDF) score will be represented for the words which have rare occurrences.

  Formula:

  $$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \quad (1)$$

  Where,

  $tf_{i,j}$ = Number of occurrences of i in j

  df_i = Number of documents containing i

  N = Total number of documents

Tools that are applied for the analysis purpose are namely; Python 3.7, NLTK toolkit for python and NLP.

3.3. Proposed framework

Data feed from four sources; Forum discussion, Twitter, News and RSS feed are used as input. On these text data pre-processing is done where removal of punctuation and stopwords is done also tokenization and lemmatization is done. Henceforth, pre-processed data is used on which three NLP techniques namely; Bag of words, TF-IDF and N-grams are applied. The feature set is created under every technique on which six ML algorithms are applied to test the sentiment prediction capability of each algorithm if it is correctly classifying sentiment under any one of the categories which are; Positive, Negative and Neutral. The flow chart of the proposed framework of sentiment analysis and prediction is depicted in Figure-3 as below.

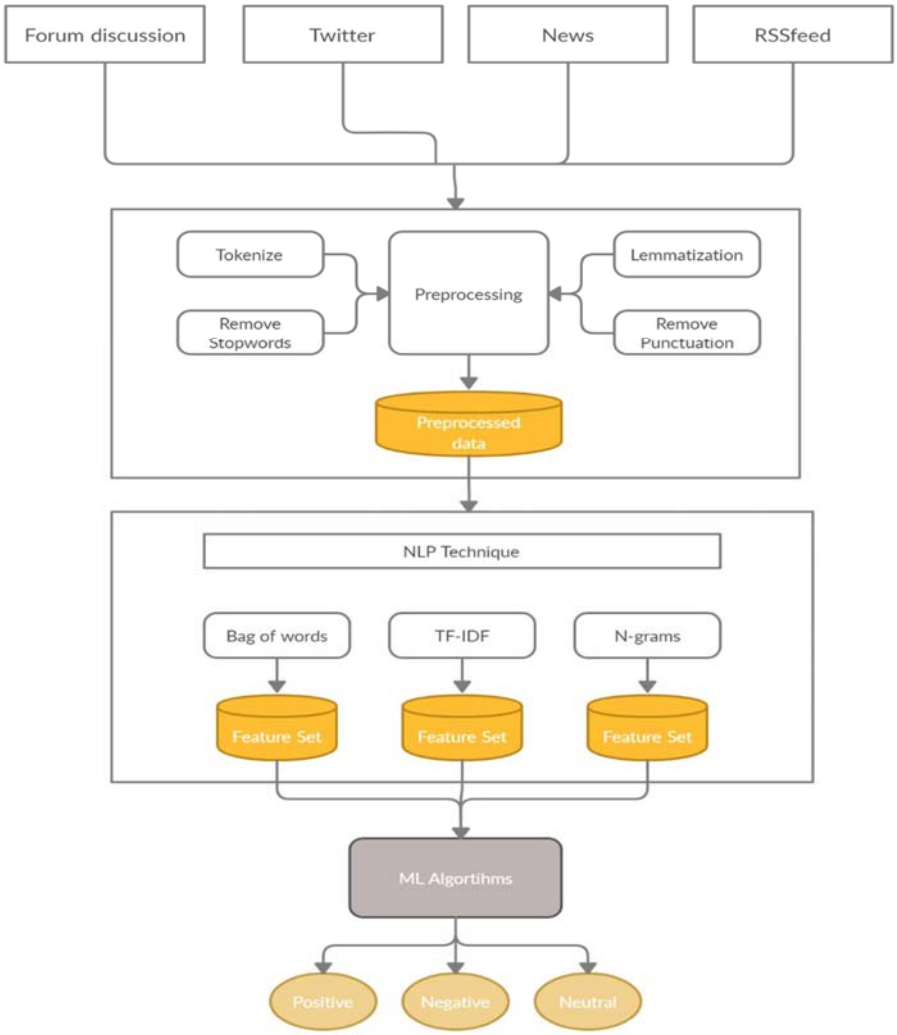

Figure 1. Proposed framework of sentiment analysis and prediction

4. Result and discussion

This study is an endeavour to identify the optimal method of sentiment analysis and sentiment prediction in context to Indian stock market news by making a comparison among various methods with specific reference to time period which covers the Covid-19 pandemic impact. For sentiment analysis and to test prediction power of algorithms for sentiment, the text is labeled under any of three categories; positive, negative and neutral. This study has tested how effectively these six algorithms with two techniques predict the sentiments by accurately labeling the text under the correct category. Figure-4 represents the accuracy level of six algorithms for sentiment prediction applied with the Bag-of-words technique as well as the TF-IDF technique.

Table 4. Classification accuracy for six algorithms.

Sr. No.	Algorithm	Technique	Classification Accuracy (%)
1	Decision Tree	Bag-of-words	77
		TF-IDF	75
2	KNN	Bag-of-words	53
		TF-IDF	68
3	Logistic Regression	Bag-of-words	78
		TF-IDF	76
4	Naïve Bayes	Bag-of-words	62
		TF-IDF	61
5	Random Forest	Bag-of-words	76
		TF-IDF	74
6	Support vector machine	Bag-of-words	78
		TF-IDF	76

While comparing Bag-of words technique with TF-IDF technique for the given data, it was observed that the Bag-of-words technique has shown higher classification accuracy in terms of sentiment prediction for five of six algorithms such as Decision Tree, Logistic Regression, Naïve Bayes, Random Forest and Support vector machine. The TF-IDF technique has shown higher accuracy only in the case of the KNN algorithm. When a comparison is made among six algorithms where Bag-of words technique is applied, both Logistic Regression and Support vector machine algorithms have indicated the highest classification accuracy of 78% for sentiment prediction. Naïve Bayes algorithm has generated the lowest classification accuracy of 62% for sentiment prediction. Below Figure-5, indicates graphically the accuracy level of all six algorithms with Bag-of words and TF-IDF technique.

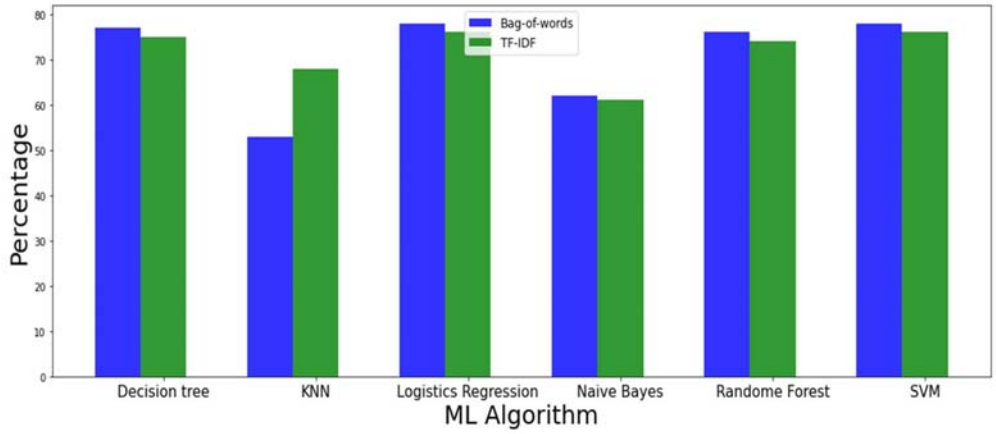

Figure 2. Bar chart representation of accuracy of six algorithms with two techniques

The recorded sentiment of the duration from 1st-January-2020 till 24th-August-2020 is depicted in Figure-6. It can be observed from the Figure-6 how the sentiment movements are gyrating.

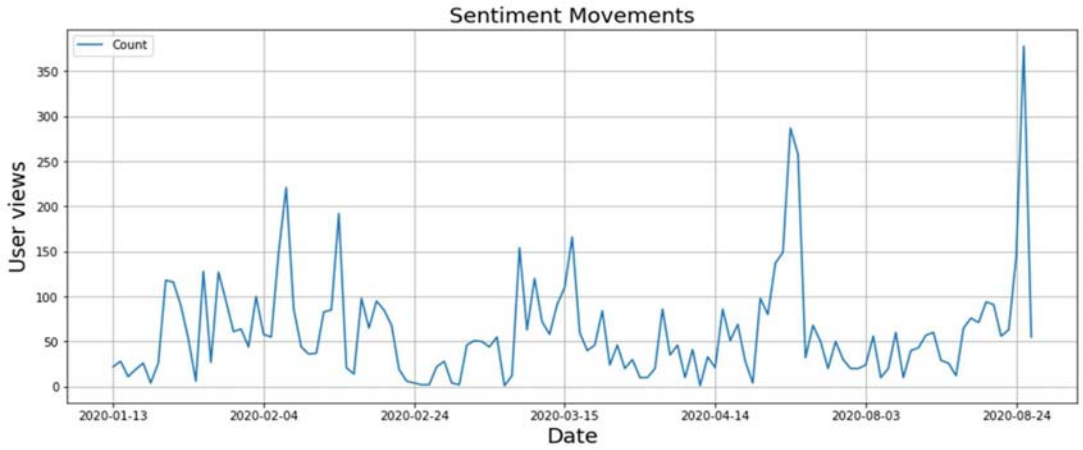

Figure 3. Sentiment graph

5. Concluding remarks and future scope

Based on the data which represents very recent time period and crucially includes the outbreak and spread of Covid-19 pandemic in India for performing the sentiment analysis to identify the optimal algorithm, Logistic Regression and Support vector machine algorithms have presented better results – making these two algorithms superior in sentiment prediction in comparison with other algorithms when Bag-of words technique is applied. The study contributes significantly towards the identification of superior algorithms in terms of accuracy for sentiment prediction. As natural language processing is increasingly becoming a potent tool for understanding text messages, this article provides a solid platform to test six selected algorithms with special reference to Indian stock markets news and covering the most recent time period of Covid-19 pandemic impact.

In the future, such research can be expanded to include other methods of sentiment analysis with more text data. Building the prediction model for stock market price movement based on identified superior algorithms in this study will throw light on if such predictive models generate better results or not. Conducting research with wider data sources and for a longer time period may also provide insight into if results generated in this study get substantiated in other such studies or not.

6. References

[1] Nasukawa, T., & Yi, J. (2003). Sentiment analysis. Proceedings of the International Conference on Knowledge Capture - K-CAP '03. doi:10.1145/945645.945658

[2] Chakraborty, K., Bhattacharyya, S., Bag, R., & Hassanien, A. A. (2019). Sentiment Analysis on a Set of Movie Reviews Using Deep Learning Techniques. Social Network Analytics, 127-147. doi:10.1016/b978-0-12-815458-8.00007-4

[3] Yadav, R., Kumar, A. V., & Kumar, A. (2019). News-based supervised sentiment analysis for prediction of futures buying behaviour. IIMB Management Review, 31(2), 157-166. doi:10.1016/j.iimb.2019.03.006

[4] Paredes-Valverde, M. A., Colomo-Palacios, R., Salas-Zárate, M. D., & Valencia-García, R. (2017). Sentiment Analysis in Spanish for Improvement of Products and Services: A Deep Learning Approach. Scientific Programming, 2017, 1-6. doi:10.1155/2017/1329281

[5] Alqaryouti, O., Siyam, N., Monem, A. A., & Shaalan, K. (2020). Aspect-based sentiment analysis using smart government review data. Applied Computing and Informatics, Ahead-of-print(Ahead-of-print). doi:10.1016/j.aci.2019.11.003

[6] Bouteska, A. (2019). The effect of investor sentiment on market reactions to financial earnings restatements: Lessons from the United States. Journal of Behavioral and Experimental Finance, 24, 100241. doi:10.1016/j.jbef.2019.100241

[7] Broadstock, D. C., & Zhang, D. (2019). Social-media and intraday stock returns: The pricing power of sentiment. Finance Research Letters, 30, 116-123. doi:10.1016/j.frl.2019.03.030

[8] Groß-Klußmann, A., König, S., & Ebner, M. (2019). Buzzwords build momentum: Global financial Twitter sentiment and the aggregate stock market. Expert Systems with Applications, 136, 171-186. doi:10.1016/j.eswa.2019.06.027

[9] S., S., & K.v., P. (2020). Sentiment analysis of malayalam tweets using machine learning techniques. ICT Express. doi:10.1016/j.icte.2020.04.003

[10] Kim, Y., & Shin, H. (2018). A New Approach for Measuring Sentiment Orientation based on Multi-Dimensional Vector Space. ArXiv, abs/1801.00254.

[11] Yadav, A., Jha, C. K., Sharan, A., & Vaish, V. (2020). Sentiment analysis of financial news using unsupervised approach. Procedia Computer Science, 167, 589-598. doi:10.1016/j.procs.2020.03.325

[12] Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review, 27, 16-32. doi:10.1016/j.cosrev.2017.10.002

[13] Bhardwaj, A., Narayan, Y., Vanraj, Pawan, & Dutta, M. (2015). Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty. Procedia Computer Science, 70, 85-91. doi:10.1016/j.procs.2015.10.043

[14] Suman, N., Gupta, P. K., & Sharma, P. (2017). Analysis of Stock Price Flow Based on Social Media Sentiments. 2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS). doi:10.1109/icngcis.2017.34

[15] Derakhshan, A., & Beigy, H. (2019). Sentiment analysis on stock social media for stock price movement prediction. Engineering Applications of Artificial Intelligence, 85, 569-578. doi:10.1016/j.engappai.2019.07.002

[16] Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Systems with Applications, 73, 125-144. doi:10.1016/j.eswa.2016.12.036

[17] Wang, Z., Ho, S., & Lin, Z. (2018). Stock Market Prediction Analysis by Incorporating Social and News Opinion and Sentiment. 2018 IEEE International Conference on Data Mining Workshops (ICDMW). doi:10.1109/icdmw.2018.00195

[18] Urolagin, S. (2017). Text Mining of Tweet for Sentiment Classification and Association with Stock Prices. 2017 International Conference on Computer and Applications (ICCA). doi:10.1109/comapp.2017.8079788

[19] Wang, Y., & Wang, Y. (2016). Using social media mining technology to assist in price prediction of stock market. 2016 IEEE International Conference on Big Data Analysis (ICBDA). doi:10.1109/icbda.2016.7509794