Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 167 (2020) 1821–1828

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2019)

Analysis of Political Sentiment Orientations on Twitter

Mohd Zeeshan Ansari[a], M. B. Aziz[a,*], M. O. Siddiqui[b], H. Mehra[a], K. P. Singh[a]

[a]*Department of Computer Engineering, Jamia Millia Islamia, New Delhi, India*
[b]*Department of Electrical Engineering, Jamia Millia Islamia, New Delhi, India*

Abstract

The dramatic increase in the number of users on social media platform leads to the generation of huge amount of unstructured text in the form of messages, chats, posts and blogs. Besides the exchange of information, social media is a remarkably convenient medium to express the ideas and opinions which gain popularity when liked by a large set of users. This popularity may reflect the sentiment of people towards that person, organization or a place. The social media platform, such as Twitter, generates huge amounts of the text containing political insights, which can be mined to analyze the people's opinion and predict the future trends in the elections. In this work, an attempt is made to the mine tweets, capture the political sentiments from it and model it as a supervised learning problem. The extraction of tweets pertaining to the General Elections of India in 2019 is carried out along with the study of sentiments among Twitter users towards the major national political parties participating in the electoral process. Subsequently, the classification model based on sentiments is prepared to predict the inclination of tweets to infer the results of the elections. The Long Short Term Memory (LSTM) is employed to prepare the classification model and compare it with the classical machine learning models.

Keywords: Political Orientations; Opinion Mining; Text Classification; Twitter

1. Introduction

Twitter is a popular social media platform extensively used for networking and microblogging where users post messages in the form of *tweets* that can be maximum 280 characters long. As of 2018, Twitter has more than 321 million monthly active users, of which 34.4 million were from India[1]. As a result, the posts generated are on a variety of topics, with news being one of the major ones, moreover, Twitter was the largest source of breaking news during the U.S. Presidential Elections in 2016. Twitter has a huge impact on the political sphere in India as the success of a leader always depends heavily on their ability to communicate with the masses. In recent times, Twitter

* Mohd Bilal Aziz. Tel.: +91-9680790209.
E-mail address: bilal.aziz9680@gmail.com

10.1016/j.procs.2020.03.201

has been extensively used as a medium to express political thoughts and connect with the common man. Although, the number of Twitter users compared to the total population of India is very small, the educated class, politicians, actors and other celebrities are a part of these numbers. These 'influencers' are capable of affecting the sentiments and opinions of their followers based on their tweets. Being able to judge sentiment would be helpful in the organization of information into groups and make it easier for users to find and react to similar or opposite opinions thereby providing a better way to discuss and share information.

This work presents the investigation of the different political preferences a user can have about the Indian electoral parties of 2019. The identification of real time political preferences over social media platforms, e.g. twitter is considerably useful in political campaigns. Such preferences may be tracked based on different regions by the analysis of data acquired, and subsequently, necessary action may be taken by the regional political affiliations. Taking into account the eight different sentiment classes based on different political alignments, the tweets are automatically extracted, cleaned and manually annotated according to annotation guidelines. The tweets were collected during a period of three months from Jan to March 2019 as this was the peak time when sentiments and opinions were at an all-time high, since it was the time just preceding the 2019 Lok Sabha elections, which were held in April. However, it must be drawn to the attention that the dataset used is extremely small when compared to the population size of the country and hence our result may not be completely indicative of the outcome. Different features are extracted and various learning approaches are examined on the preprocessed data. The Long Term Short Memory and various machine learning models such as Support Vector Machines, Decision Tree, Logistic Regression and Random Forest are being considered for the investigation.

The overall contribution is organized in multiple sections as follows. Section 2 examines various state-of-the-art political sentiment analysis and social media mining. Section 3 introduces the mathematical framework for the political preferences of Twitter users with respect to political parties. Section 4 elaborates the methodology of data preparation which includes extraction of raw tweets from twitter, preprocessing and normalization of data, manual annotation, and text representation in different forms. This section also introduces conventional classification methods suitable for the analysis of data acquired. Section 5 investigates the corpus statistics, various experiments conducted over different configurations and finally puts forward the comparative analysis of the results obtained. In Section 6, we discuss the limitations of the presented work and the scope for future work.

2. Previous Work in Sentiment Analysis and Political Orientations

Sentiment analysis is generally done in a two-phase format, where in the first phase, the relevant data are collected, and in the second phase, the extraction of the sentiment takes place. Tweets may be considered relevant if they contain words from a list of target keywords that is compiled either manually, such as, in the case of Wang et. al. (2012), O'Connor et. al. (2010) and Tumasjan et. al. (2010) [1-3], or in a semi-automatic fashion as in the case of (Conover et. al. 2011a, b)[4,5] through the expansion of a seed set. Subsequently, after the compilation of pertinent tweets, various approaches are employed to extract the sentiment from the tweet. Unsupervised methods rely on lexicons, a list of 'positive' and 'negative' keywords, estimating a sentiment based on the ratio of occurrence of two types of keywords with respect to one another, or just by counting the frequency of each term [2,6]. A similar study that made use of lexicons was conducted by Mishra et. al. (2016) where public sentiment with regards to the 'Digital India' mission was analyzed based on the ratio of positive to negative keywords [7]. More advanced approaches make use of supervised learning techniques, and prediction models are trained on either tweets that have been manually classified [1, 4], or on tweets with an emotional context (Marchetti-Bowick and Chambers 2012) [8], i.e., hashtags and emojis, such as :D, #upset, #excited, etc. Research studies by A. Jain and P. Dandannavar (2017) focused on combining a lexical based approach with a learning based approach to form a hybrid approach to sentiment analysis [9]. Social participation by users can be accredited with the rise of citizen participation in the electoral processes. Such participation occurs not only during the lead up to the main day, i.e. during the campaign, but also on Election Day itself. According to Wagner and Gainous (2013), this participation brings forth a new environment where active participation on social media may increase the citizen's role in politics and also empower the democratic process on different levels [10]. Research in social media mining and network analysis has generated interest in analyzing the large amounts of data that is produced worldwide. This data is useful when accumulated

[1]https://www.statista.com/statistics/381832/twitter-users-india/

and examined in order to discover behavioral patterns and interactions, and also obtain a better understanding of unrelated issues [11]. This method was also utilized by Prati and Said-Hung (2017) where they studied the exchange of messages that had a "defined ideological load", as well as the citizen participation on Twitter during the Spanish elections during May 2015 [12]. The analysis of separation levels is also carried out to observe the political orientations in tweets and comments along with the consideration of the event timelines [13,14]. Other studies, such as the ones done by Koc-Michalska et. al. (2014) and Fominya (2014), focused on conventional sources of media, social movements, electoral campaigns and voter engagement [15,16]. Also, the study put forward by Conover et. al. (2011) examined the partisanship of voters with respect to political leanings and how the two groups interact with each other in the form of retweets or mentions [4]. Sahayak et. al. (2015) focused on sentiments pertaining to a variable search query, with the intention of making it easier for companies to gather feedback about their products, and for customers who want to search the opinions of others regarding a product, before purchase [17].

3. Background

Given a list of parties $Pi \in P$, P_0 stands for the Bharatiya Janta Party (BJP), P_1 stands for Indian National Congress (INC), P_2 stands for Other parties. Also, given a user *u* and a tweet *t*, say *u* prefers p if the user has at least one tweet that contains a term from the preference profile of p. It's reasonable to assume that during a campaign, parties will use Twitter to advertise their candidates, propose party policies, participate in debates and interviews, and also criticize their opponents. The aim is to utilize the content generated from such behaviour to capture party-specific topics in the form of unigrams, bigrams or trigrams. Given a set *P* of parties, associate with each party $Pi \in P$, a document *dp* that is modelled as a *bag of words*. Let D = { *dp* | $Pi \in P$ }. Term probabilities in the corpus are calculated using *tf-idf* scores.

4. Methodology

4.1 Data Collection

The collected 3896 tweets are prominently significant to the two main national parties: say P_0, P_1 and the rest of the parties as Others, say P_2. The relevant tweets are identified by querying Twitter for party names, their abbreviations, and by the names of their most popular leaders along with the election hashtag, #LokSabhaElections, #ElectionsInIndia, etc. For this study and analysis, only two major parties are taken into consideration and the rest of the parties are merged into Others, over which the analysis is performed collective in nature. This is in contrast to measure the sentiment of each party in isolation which may poorly allow the classifier to discriminate between the parties. Therefore, it is not adopted that the distribution of sentiments reflected in tweets significant towards a single party is suggestive of sentiment to that party. Subsequently, in order to incorporate the inclination of tweets towards or against the political parties collectively, it is reasonable to measure the combined effect of the parties. To allow this, the considered class of sentiments include those that also reflect P_0 and P_1, P_0 and P_2, P_1 and P_2, none and unrelated. Significantly, the mining of tweets from twitter was conducted from Jan to March 2019.

Finally, the fields selected are (1) ID: A unique ID given to each tweet. The use of a numerical ID is preferred and not the username as an identifier in order to keep the users anonymous. The main idea is to get the number of people belonging to any of the camps. (2) Text: Shows the original tweet posted by the user. (3) Class: Three columns, two of them belonging to the two major parties (i.e. BJP and Congress) and the last column for any other party. These columns accept binary data based on whether the tweet is aligned towards any one, two or all of them or even none of them at all. The 3 columns are taken together, and the binary number formed is converted to decimal to serve as an easy classifier. It can be from the range of 0-7.

4.2 Preprocessing

Twitter has distinct conventions as compared to other online social media text due to which special care is taken during cleaning and pre-processing of raw tweets. The pre-processing includes the following steps: (1) non-standard lexical tokens such as mentions, hashtags, emoticons, unconventional punctuation, are filtered out just after tokenization, (2) duplicate tweets and retweets are eliminated in order to maintain the uniqueness of each tweet, (3) standard *stopwords* are removed, (4) finally, *case folding* is performed to convert all the tokens into lower case letters. A fair number of out of vocabulary (OOV) words are observed in the data which are kept intact.

4.3 Annotation

Three annotators potentially conscious with the electoral politics and sensibly aware with the keywords used in favour or against a party or a leader were involved for annotation. Annotators were instructed to identify the opinion or speculation present in each tweet instead of an absolute positive or negative sentiment. The annotation categories consisted of eight classes which are as follows: P_0, P_1, P_2, (P_0 or P_1), (P_0 or P_2), (P_1 or P_2), *none* and *non-relevant*. Instead of discarding the group of irrelevant annotations, it was assigned a proper class to identify the distribution of political and non-political tweets. The following definition of sentiment is proposed in (Wilson et al, 2005) [18]. The manually labelled corpus achieved the inter-annotator agreement over 98 %.

4.4 Feature extraction

Four different types of feature vectors taken into consideration are based on Term-Frequency; Term Frequency–Inverse Document Frequency for unigrams, bigrams and trigrams (denoted as *tf*, *tf-idf1*, *tf-idf2*, *tf-idf3* respectively). The Term Frequency – Inverse Document Frequency (TF-IDF) is a significant statistical measure used to convert a collection of raw documents to a matrix in the form of term frequency vs inverse document frequency. In other words, it assigns weightage to words on the basis of rarity in a document. TF-IDF formula is expressed as,
for a term *i* in document *j*:

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

where $tf_{i,j}$ = number of occurrences of i in j, df_i = number of documents containing *i*, *N* = total number of documents.

4.5 Machine learning classification algorithms

Support Vector Machines (SVM). A discriminative classifier, also known as the Support Vector Classifier (SVC) is a machine learning algorithm derived from statistical learning theory based on the principle of structural risk minimization. Introduced by Vapnik (1995) [19], it aims to reduce test error and computational complexity. The SVM identifies a separating hyper-plane with optimality between the two classes of a training dataset. This optimum separation is obtained by the hyper-plane that has the largest distance from the nearest training dataset. SVM are conventionally used as supervised learning methods used for binary classification, regression and outlier detection.

Decision Tree Classifier (DTC). A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. It performs recursive binary partition of the feature space to break down a data set into smaller and smaller subsets while simultaneously developing an incremental decision tree. This results in a tree with decision and leaf nodes. A decision node has two or more branches; a leaf node represents decisions or classifications. The topmost node is referred to as the root node. The smallest tree that fits the data is then found. Usually, this is the tree that yields the lowest cross-validation error. The DTC was selected because of its ability to empower predictive models with high stability, ease of interpretation and accuracy.

Table 1. Corpus Statistics

Sentiment Classes	No. of tweets in each class	% of tweets in each class	No. of tokens in each class	% of tokens in each class
None	5	0.12	60	0.16
P_0	2161	**55.46**	20453	57.01
P_1	515	13.21	3792	10.57
P_2	81	2.07	842	2.34
P_0 and P_1	13	0.33	151	0.42
P_0 and P_2	355	9.11	3434	9.57
P_1 and P_2	666	17.09	6153	17.15
Non relevant	100	2.56	988	2.75
Total	3896	100	35873	100

Logistic Regression (LR). A common classification method for many applications, such as text classification, computer vision, etc. calculates the probability of a sample that belongs to a particular class by learning the set of parameters that minimize the induced log-likelihood over the training examples.

Random Forest Classifier (RFC). A random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(x, \theta_k), k = 1,...\}$ where the $\{\theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x. (L. Breiman, 2001) [20] The Random Forest is based on the bagging algorithm and uses the Ensemble learning technique. It creates many random trees on the subset of data and combines the output of all trees. The advantage of an RFC is that it gives more accurate predictions when compared to simple Decision Tree models, by reducing overfitting and variance.

Long Short Term Memory (LSTM). Long short-term memory (LSTM) units are units of a recurrent neural network (RNN) that are capable of learning long-term dependencies. An LSTM network is basically an RNN composed of LSTM units. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM is a type of deep learning model that is mostly used for the analysis of sequential data (time-series data prediction). It is applied to process tasks such as speech recognition, music composition, language translation etc. [21]

Table 2. Performance of classifiers over various features.

Classification method	Precision				Recall				F1-Score			
	tf	*tf-idf1*	*tf-idf2*	*tf-idf3*	*tf*	*tf-idf1*	*tf-idf2*	*tf-idf3*	*tf*	*tf-idf1*	*tf-idf2*	*tf-idf3*
Support Vector Machines	0.38	0.31	0.31	0.31	0.57	0.55	0.55	0.55	0.43	0.39	0.39	0.39
Decision Tree Classifier	0.71	0.69	0.68	0.69	0.71	0.71	0.68	0.71	0.71	0.70	0.67	0.70
Logistic Regression	0.70	0.71	0.72	0.70	0.70	0.73	0.73	0.72	0.70	0.69	0.69	0.68
Random Forest Classifier	0.73	**0.77**	0.74	0.76	0.76	0.76	0.75	**0.77**	0.73	**0.74**	0.72	**0.74**
Long Short Term Memory	0.72	0.73	0.70	**0.76**	0.73	0.72	0.70	0.75	0.71	0.70	0.69	**0.74**

Table 3. Class-wise Performance of LSTM using *tf-idf3*

Classes	Precision	Recall	F1 Score
None	0.88	0.7	0.78
P_0	0.78	0.92	0.85
P_1	0.73	0.77	0.75
P_2	0.33	0.12	0.18
P_0 and P_1	0	0	0
P_0 and P_2	0.81	0.61	0.7
P_1 and P_2	0.72	0.39	0.5
Non-relevant	0	0	0

5. Experiments and Results

The analysis of annotated corpus generates the statistics as shown in Table 1. Among the 3896 collected tweets, a majority of 55.46 % tweets show the sentiments towards the party P_0, which includes 20453 tokens. In contrast to this, the percentage of tweets that show sentiments towards P_1 is only 13.21 % and towards P_2 is a drastically lesser percentage of 2.07 % only. This shows that the appearance of users on Twitter favouring the party P_0 is quite large. Interestingly, the tweets that show favour in a combination of parties P_0 and P_1 is extremely low of 13 tweets which is 0.33 % of the total tweets. However, the favour observed for P_0 and P_2 is 9.11 % and P_1 and P_2 is 17.09 % which is substantially large. This indicates that there are negligible users that are of opinion for both P_0 and P_1.

Besides LSTM, to comparatively evaluate the corpus, various machine learning algorithms such as Support Vector Machines, Decision Tree, Logistic Regression, and Random Forest are employed (see Table 2). For validation, stratified 10-folds are cross-validated over the whole set of data. The Random Forest with *tf-idf-unigram* and LSTM with *tf-idf-trigram* exhibit a precision of 0.77 and 0.76 which is highest in all the classifiers. The recall of 0.77 is highest for Random forest with *tf-idf-trigram* which is comparable with its *term-frequency* and *tf-idf-unigram* models. The Random Forest and LSTM show the highest F1-Score of 0.74 for *tf-idf-unigram* and *tf-idf-trigram* models. In the overall analysis, the performance of Support Vector Machine is poorest among all. Fig. 1 shows the comparative picture of the accuracy with respect to all the classifiers. The highest accuracy is achieved by the LSTM classifier when applied to *tf-idf-trigram* method for feature selection.

Table 3 presents the class-wise precision, recall and F1-Score for all eight classes. Among the relevant classes, the highest precision is obtained is 0.81 for class P_0 and P_2 and the lowest precision obtained is 0 for class P_0 and P_1. Remarkably, the highest recall is obtained for class P_0 and the highest F1-Score obtained is also for class P_0. The *None* class obtaining significant values of performance measures may be due to the clear discrimination with all the other classes as the other classes contain the sentiments but *None* class does not contain any sentiment. The zero values in case of P_0 and P_1, and *Non-relevant* class is due to fairly small number of training examples in these classes.

Table 4. Classifier training time using *tf-idf3*

Classifier	Training time (s)
Support Vector Machines	4.4699
Decision Tree Classifier	0.1215
Logistic Regression	**0.0519**
Random Forest Classifier	0.8587
Long Short Term Memory	82.7816

Fig. 1. Comparison of accuracies for classifiers on different features

It is also useful to have an idea of the most common words floating around in these tweets. This information is shown in the form of a word cloud in Fig 2. The execution time of training the algorithms is recorded and presented in Table 4. The Logistic Regression Classifier takes the minimum time of 0.0519 seconds whereas LSTM takes an extremely large maximum time of 82.7816 seconds. SVM takes a comparatively large amount of time, with respect to Decision Tree, Logistic Regression and Random Forest. Among all, Random Forest takes the optimally low time of 0.8587 seconds for training along with high performance on the data as evaluated earlier. It can be concluded that Random Forest performs best along with reasonably optimal training time, however, LSTM performs slightly better in terms of accuracy when compared to Random Forest, but at a high computation cost.

6. Conclusion

The preparation of the annotated corpus from Twitter automatically for political sentiment analysis of national political parties of India with respect to the forthcoming general elections in India is carried. Inspection of the sentiment keywords pertaining to political preferences in the tweets is considered during annotation in order to mark class labels. Several significant features based on term frequency inverse document frequency are extracted from the corpus to examine the various learning models. The data analysis of the acquired corpus reflects the presence of a single party on Twitter which is more than fifty percent among all the extracted politically motivated tweets. The evaluation of the annotated corpus over itself using LSTM and several machine learning models are obtained and promising results are achieved with LSTM and Random Forests. For the improvement of such kind of analysis and inference, number of tweets extracted must be increased to a fairly reasonable size. The proper sampling methods must be incorporated in order to balance the class distribution among all the classes. Semi-supervised techniques for corpus construction may be employed to build a strong corpus.

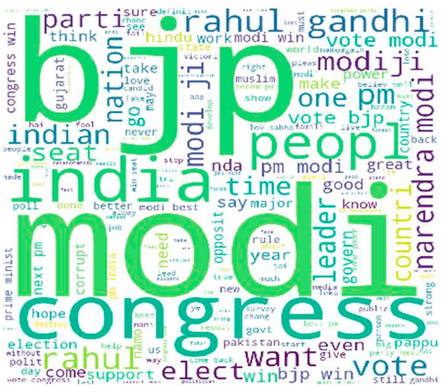

Fig. 2. Word cloud of corpus collected

References

[1] Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S (2012) "A system for real-time Twitter sentiment analysis of 2012 US Presidential election cycle." In ACL (System Demonstrations), pp. 115–120

[2] O'Connor B, Balasubramanyan R, Routledge BR, Smith NA. (2010) "From tweets to polls: linking text sentiment to public opinion time Series." In Proceedings of the ICWSM Conference.

[3] Tumasjan A, Sprenger TO, Sandner PG, Welpe IM. (2010) "Predicting elections with Twitter: What 140 characters reveal about political sentiment." In Proceedings of the ICWSM Conference.

[4] Conover M, Gonc¸alves B, Ratkiewicz J, Flammini A, Menczer F. (2011a) "Predicting the political alignment of Twitter users." In Proceedings of SocialCom/PASSAT Conference, pp. 192–199

[5] Conover M, Ratkiewicz, J, Francisco M, Goncalves B, Menczer F, Flammini A. (2011b) "Political Polarization on Twitter." In International AAAI Conference on Web and Social Media.

[6] Choy M, Cheong MLF, Laik MN, Shung KP (2011) "A sentiment analysis of Singapore presidential election 2011 using twitter data with Census correction." CoRR abs/1108.5520

[7] Mishra P, Rajnish R and Pankaj Kumar (2016) "Sentiment Analysis of Twitter Data: Case Study on Digital India." INCITE-2016. Amity University, India.

[8] Marchetti-Bowick M, Chambers N. (2012) "Learning for microblogs with distant supervision: political forecasting with Twitter." In Proceedings of the EACL Conference, pp. 603–612

[9] Jain, Prakash, and Padma Dandannavar (2017) "Text Analytics Framework Using Apache Spark and Combination of Lexical and Machine Learning Techniques." *International Journal of Business Analytics and Intelligence* **5** (**1**): 36–42.

[10] Wagner, Kevin, and Jason Gainous (2013) "Digital uprising: The Internet Revolution in the Middle East." *Journal of Information Technology and Politics* **10**(3): 261–275

[11] Xu G, Li L (2013). "Social media mining and social network analysis: emerging research, information." Science reference. IGI Global, Hershey.

[12] Prati RC, Said-Hung E (2017) "Predicting the ideological orientation during the Spanish 24M elections in Twitter using machine learning." doi 10.1007/s00146-017-0761-0

[13] Gruzd, Anatoliy, and Jeffrey Roy (2014) "Investigating Political Polarization on Twitter: A Canadian Perspective" *Policy Internet* **6**(2)

[14] Elmer, Greg (2012) "Live research: Twittering an election debate." *New Media & Society* **15**(**1**): 18–30

[15] Koc-Michalska, Karolina, Darren G. Lilleker, Pawel Surowiec, and Pawel Baranowski (2014) "Poland's 2011 Online Election Campaign: New Tools, New Professionalism, New Ways to Win Votes" *Journal of Information Technology and Politics* **11**(**2**): 186–205

[16] Fominaya CF (2014) "Social movements and globalization: how protests, occupations and uprisings are changing the world." Palgrave Macmillan, New York.

[17] Sahayak, Varsha, Vijaya Shete, and Apashabi Pathan (2015) "Sentiment Analysis on Twitter Data." *International Journal of Innovative Research in Advanced Engineering* **2**(**1**): 178–183.

[18] Wilson, Wiebe, and Hoffmann (2005) "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 347–354.

[19] C. Cortes, V. Vapnik, "Support vector machines", *Machine Learning*, pp. 273–297 (1995).

[20] Breiman, Leo (2001) "Random Forests" *Machine Learning* **45**(**1**): 5–32

[21] Hochreiter Sepp, Jrgen Schmidhuber (1997) "Long short-term memory", *Neural computation*, **9 (8)**, pp. 1735-1780