Determining the Popularity of Political Parties Using Twitter Sentiment Analysis

Sujeet Sharma and Nisha P. Shetty

Abstract With the advancement in the Internet Technology, many people have started connecting to social networking websites and are using these microblogging websites to publically share their views on various issues such as politics, celebrity, or services like e-commerce. Twitter is one of those very popular microblogging website having 328 million of users around the world who posts 500 million of tweets per day to share their views. These tweets are rich source of opinionated User-Generated Content (UGC) that can be used for effective studies and can produce beneficial results. In this research, we have done Sentiment Analysis (SA) or Opinion Mining (OM) on user-generated tweets to get the reviews about major political parties and then used three algorithms, Support Vector Machine (SVM), Naïve Bayes Classifier, and k-Nearest Neighbor (k-NN), to determine the polarity of the tweet as positive, neutral, or negative, and finally based on these polarities we made a prediction of which party is likely to perform more better in the upcoming election.

Keywords Sentiment analysis • Opinion mining • Tokenization
Classification • Natural language processing (NLP)

1 Introduction

The rapid development in the Internet technology has also accelerated the usage of microblogging websites. There are a lot of such websites like Twitter, Facebook, Tumblr, etc. Using these services, people not only gets connected to their family and friends but also uses these services for sharing their views or opinion publically

S. Sharma (✉)
Sikkim Manipal Institute of Technology, Majitar, East Sikkim 737136, India
e-mail: sujeetsharma1107@gmail.com

N. P. Shetty
Manipal Institute of Technology, Manipal University, Manipal 576104, India
e-mail: nisha.pshetty@manipal.edu

S. C. Satapathy et al. (eds.), *Information and Decision Sciences*,
Advances in Intelligent Systems and Computing 701,
https://doi.org/10.1007/978-981-10-7563-6_3

about various issues like any product, celebrity, or politics. During the election time, these microblogging websites get full of comments and reviews about the political parties and their leaders. Such reviews or comments are known as User-Generated Content (UGC) [1].

These UGCs are rich source of opinionated data and performing opinion mining on it can provide some beneficial results. "Politics using social media" [2] is now becoming a popular research area. Opinion mining is a kind of Natural Language Processing (NLP) used for identifying the mood of the public about a particular person, product, or service. In this research, opinion mining is performed on the dataset which is downloaded from one of the popular microblogging website Twitter [3] to compare the accuracy of classifiers in determining the popularity of various political parties by finding the solution to the following mentioned problems:

1. Classification of the downloaded tweets based on which political party or leader they belong to.
2. Determining the polarity of the tweets as positive, neutral, or negative.
3. Conclusion based on the overall polarities of the tweets for individual parties or leaders.

For example, "So proud of Corbyn, thanks for representing us. What a man-I#election#myvote4labour". Here, "Corbyn" is the leader of the Labour party in the United Kingdom General Election 2017, "proud of Corbyn" and "myvote4labour" make this tweet's polarity as positive.

The opinionated words (usually adjectives) and important hashtags like (#myvote4labour) are first extracted and classified as positive, negative, and neutral using the appropriate library and packages. Based on that, the polarity of the whole tweet is determined and finally summarization of all the tweets can be used for decision-making of which party or leader is more likely to perform better in the upcoming election. Opinion mining in this research is performed using R Programming Language [4].

The sections in this paper are categorized in the following manner. Section 1 describes the need and advantages of the opinion mining or sentiment analysis. Section 2 describes the related work performed by distinguished researchers in the field of opinion mining and sentiment analysis. Section 3 describes the proposed methodology and workflow. Section 4 describes the obtained results. The last section describes the comparison and conclusion of the obtained results along with the future work.

2 Related Work

Shengyi et al. [5] have introduced an improved version of k-NN algorithm named INNTC for text categorization. They have used one-pass clustering algorithm with approximately linear time complexity which is an incremental clustering algorithm capable of handling large and high-dimensional data. After forming the clusters, they have used the k-NN classifier for the text categorization. And finally, they compared the accuracy and showed that their algorithm performs better than SVM in many of the datasets.

Karim et al. [6] have worked in two very popular data mining techniques: the Naïve Bayes and the C4.5 decision tree algorithms. They have used publically available dataset UCI data for training and testing the models in order to predict the chances of whether a client will be subscribing to a term deposit or not. They got the accuracy of 93.96% for C4.5 decision tree and 84.91% for Naïve Bayes classifiers. Later in their work, they also used decision tree for extraction of interesting and important decision in business area.

Hsu et al. [7] have worked on the SVM classifier and showed when to use linear kernel and when to use RBF kernel. They proved that if the number of features is small, then data can be mapped to higher dimensional spaces by using nonlinear kernel. But in case of large number of features, one should not map data to a higher dimensional space as nonlinear mapping will not enhance the performance. So in case of large number of features, the linear kernel can perform good enough. They also proved that RBF kernel is at least as good as linear kernel (which holds true only after searching the (C, γ) space, where C and γ are two parameters for an RBF kernel.)

Kousar Nikhath et al. [8] have implemented an email classification application based on text categorization, using k-NN classification algorithm and achieved the accuracy of 74.44%. They implemented the application using two processes: training process and classification process, where the training process uses a previously categorized set of documents (training set) to train the system to understand what each category looks like. And then the classifier uses the training model to classify new incoming documents. They also used Euclidean distance as a similarity function for measuring the difference or similarity between two instances.

In [9], Alexandra Balahur and Marco Turchi have first implemented a simple sentiment analysis system for English tweets and later they extended it to multilingual platform. With the help of Google machine translation system, they translated the training data to four different languages—French, German, Italian, and Spanish. And finally they proved how overall accuracy of sentiment analysis can be improved from 64.75% (in English) to 69.09% (in all 5 languages) by the jointly using the training data from different languages.

Dilara et al. [10] have performed sentiment classification on Twitter Sentiment 140 datasets and proved that using their method Naïve Bayes can outperform the accuracy of SVM classifier also. Since Naïve Bayes has a disadvantage, whenever it is applied to high-dimensional data, i.e., for text classification it suffers from sparsity, so they deduced the smoothing as the solution of this problem. He used Laplace Smoothing in his work.

In [11], Minqing Hu and Bing Liu have proposed a technique which is based on Natural Language Processing (NLP) and Data Mining method for mining and summarizing the product reviews. They have first identified the features of the product mentioned by the users and then identified the opinionated sentence in each of the reviews present and classified it as positive or negative opinions and finally produced a conclusion from the determined information.

Chien-Liang et al. [12] have designed and developed a movie rating and review summarization system in a mobile environment, where movie rating is based on the sentiment classification results. They used feature-based summarization for the generation of condensed descriptions of the movie review. They proposed a novel approach to identify the product features which is based on latent semantic analysis (LSA). In addition to the classification's accuracy, they also considered system response time in their system design. Their work also has an added advantage where users, based on their interest, can choose the features; this module employees an LSA-based filtering mechanism, which could efficiently minimize the size of summary.

3 Methodology

Figure 1 represents the flow of data through various modules used in the proposed methodology.

3.1 Collection of Corpus

The datasets are directly downloaded from the microblogging website Twitter using the Twitter API and twitteR [13] package in R. Since we are considering the

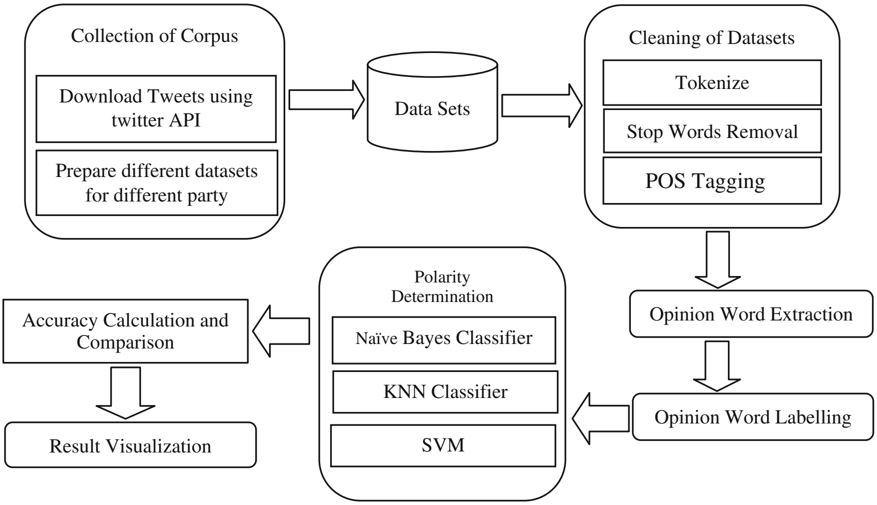

Fig. 1 Flow chart for the proposed methodology

"General Election 2017" and tweets about this election generally contains hashtags as "#GE2017" or "GeneralElection17", we have first downloaded the tweets having these hashtags. After this, the tweets are divided into different datasets based on which political party belongs to.

For example, the tweets about "Labour" party whose leader is "Jeremy Corbyn" contain words like "labour" or "corbyn", so tweets with such keywords are placed in Labour's dataset.

Then, from each of the datasets, 240 tweets are randomly selected which consists of all positive, negatives, and neutral tweets upon which opinion mining is to be performed. This method of corpus collection can be used for collecting any kind of datasets in many different languages because twitter API allows to select the languages of the tweets to be collected.

3.2 Cleaning of Datasets

The collected dataset contains a lot of ambiguous words, so as to remove those words some cleaning processes need to be done on the collected datasets. All the blank spaces and punctuation marks are removed first, then tokenization is performed in order to convert all the words into individual tokens. The next step is to remove stop words; these are most commonly used words in a language and cannot help in polarity determination. Then, the final step is to apply sentence token annotations and word token annotations to the datasets.

3.3 Collection of Opinionated Words

Extraction of opinionated words from the cleaned datasets is then performed. This step involves two process: first is to do POS tagging and second is to extract the hashtags present. POS tagging refers to assigning each of the tokens present in a corpus with its corresponding parts of speech (nouns, pronouns, adjectives, etc.). Polarity of the sentences is mostly determined by the adjective words; therefore, after performing POS tagging we can directly get the opinionated words [14]. But in the case of election, hashtags play more important role than adjective words; here public usually uses hashtags like "#myvote4conservative" to express positive feelings toward "Conservative" party, and "#toriesout" for expressing negative feeling for the same party. So along with adjective words, hashtags are also extracted to be used as opinionated words.

3.4 Labeling of Opinionated Words

The collected words are then checked upon the dictionary comprising of two files positive words and negative words, and based on that a polarity is assigned to the tweet. If there is a positive word or hashtag in the tweet, it will be assigned as +1 and for each of the negative words, −1 is assigned.

For example, “Corbyn will win#myvote4corbyn” in this tweet “win” and “myvote4corbyn” are found in positive words and hence 1 + 1 = 2, so +2 will be the score of this, and then based on the score is positive or negative, the polarity will be assigned as +1 or −1, respectively. For the words not found in dictionary, the score of 0 is assigned and tweets with score 0 will be considered as neutral with polarity as 0.

The dictionary used can be modified as per the user requirements, as in different elections and for different political parties, people use different hashtags. So accordingly user can add the new hashtags with the polarity in the dictionary.

3.5 Classification of the Tweets

Classification algorithm will be applied to the datasets so that the model can be trained on the training datasets and can be used for predicting the future trends of the data. There are many classification algorithms available; out of them, we are considering three algorithms in this paper, namely, Support Vector Machine (SVM) [15], Naïve Bayes Classifiers, and k-Nearest Neighbors (k-NN) classifiers.

4 Experiment Results

There are six political parties participating in the UK General Election 2017, out which there are only two major parties which play significant role in the election. So in this work only those two political parties, “Labour” and “Conservative”, are considered. Datasets of 240 tweets comprising all positive, neutral, and negative tweets for both of the parties are used for Opinion Mining.

4.1 Experimental Steps

1. Download the datasets for the required elections.
2. Two datasets are prepared, one for the “Labour” party and the another for the “Conservative” party.

Table 1 Classifiers with their obtained accuracy for both political parties' datasets

Classifiers	Political parties		Average accuracy (%)
	Conservative party (%)	Labour party (%)	
Naïve Bayes	64.58	61.46	63.02
k-NN	92.71	91.67	92.19
SVM	72.92	75	73.96

3. Read the datasets separately.
4. Perform the cleaning of the datasets.
5. Polarity assignment is done.
6. 60% of the datasets are considered as training sets and remaining 40% as test sets are prepared.
7. Three different classifiers are used for classification of the tweets, i.e., Naïve Bayes Algorithm [16], SVM (Support Vector Machine) [17], and k-NN (k-Nearest Neighbors) [18].
8. Calculation of accuracy for each of the classifiers.
9. Comparison of the performance of the above-stated algorithms and visualization.
10. Comparison of the performance of two parties and visualization for the prediction.

By analyzing the experimental results as shown in Table 1, it is observed that performance of k-NN classifier is more accurate than the other two classifiers. Here, k-NN is giving more accuracy as it is used similarly to the INNTC [5], where clustering is performed by collecting just the adjective and hashtags of the tweets from the training sets, which reduces the test similarity computation to a great extent and therefore reduces the impact on performance which is affected by single training sample (noisy sample) and then the k-NN algorithm is applied. Graphical representation of the same is shown in Figs. 2 and 3.

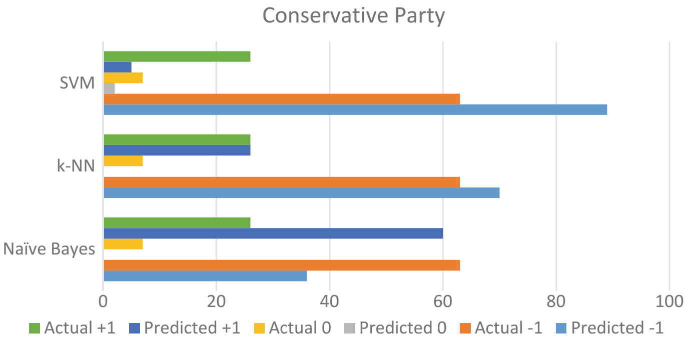

Fig. 2 Comparison of classifiers on Conservative party datasets

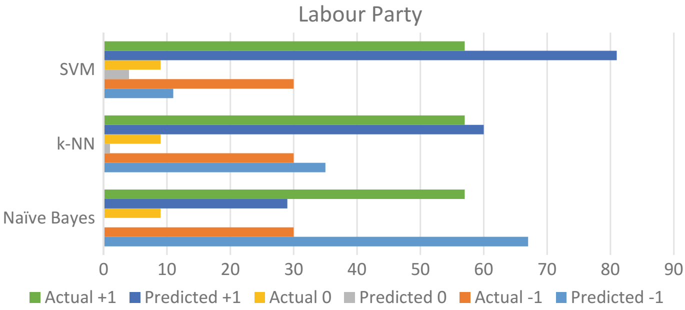

Fig. 3 Comparison of classifiers on Labour party datasets

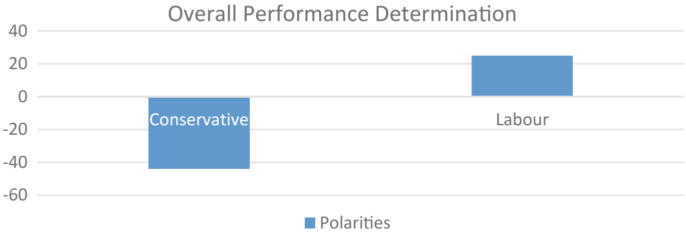

Fig. 4 Overall performance determination

The graph below is drawn for determining the future performance of both of the political parties in the upcoming elections using k-NN classifiers.

Considering the above-drawn graphs, in Fig. 4, it can be easily predicted that Conservative party's performance is likely to be decreased in the next election, whereas Labour party is predicted to perform better in the next election.

5 Conclusion and Future Work

This proposed work shows how datasets can be directly downloaded from the microblogging websites and opinion mining or sentiment analysis can be performed on it to get some useful results which can help later in decision-making. Three classifiers SVM, Naïve Bayes, and k-NN are used for opinion mining out of which k-NN gives most accurate results. Later result obtained from k-NN classifier is used to determine the performance of political parties in the future election. Future work can be performed to improve the accuracy of the predictions based on previous performance. The proposed model can be used for both election campaigning and performance prediction of the political parties. Also, it can be used by common people to know the present reputation of any political party or any politician.

References

1. Yadav, S.K.: Sentiment analysis and classification: a survey. Int. J. Advanc. Res. Comput. Sci. Manag. Studies, **3**(3) (2015)
2. Taimur, I., Ataur, R.B., Tanzila, R., Mohammad, S.U.: Filtering political sentiment in social media from textual information. In: 5th International Conference on Informatics, Electronics and Vision (2016)
3. Alexander, P., Patrick, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: International Conference on Language Resources and Evaluation (2010)
4. Akhil Kumar, K.V., Manikanth Sai, G.V., Shetty, N.P., Chetana, P., Aishwarya, B.: Aspect based sentiment analysis using R programming. In: Proceedings of Fourth International Conference on Emerging Research in Computing, Information, Communication and Applications (ERCICA-2016)
5. Shengyi, J., Guansong, P., Meiling, W., Limin, K.: An improved K-nearest-neighbour algorithm for text categorization. Exp. Syst. Appl. **391**(3) (2012)
6. Karim, M., Rahman, R.M.: Decision tree and Naïve Bayes algorithm for classification and generation of actionable knowledge for direct marketing. J. Softw. Eng. Appl. **6**, 196–206 (2013)
7. Hsu, C.W., Chang, C.C., Lin, C.J.: A Practical Guide to Support Vector Classification. www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
8. Kousar Nikhath, A., Subrahmanyam, K., Vasavi, R.: Building a K-nearest neighbour classifier for text categorization. Int. J. Comput. Sci. Informat. Technol. **7**(1), 254–256 (2016)
9. Alexandra, B., Marco, T.: Improving sentiment analysis in twitter using multilingual machine translated data. In: Recent Advances in Natural Language Processing (2013)
10. Dilara, T., Gurkan, T., Ozgun Sagturk, Ganiz, M.C.: Wikipedia based semantic smoothing for Twitter sentiment classification. IEEE (2013)
11. Minqing, H., Bing, L.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'04 pp. 168–177, ACM New York (2004)
12. Chien-Liang, L., Wen-Hoar, H., Chia-Hoang, L., Gen-Chi, L., Emery, J.: Movie rating and review summarization in mobile environment. IEEE Trans. Syst. Man Cybernet. Part C Appl. Rev. **42** (2012)
13. twitteR Package. https://cran.r-project.org/web/packages/twitterR/twitteR.pdf
14. Pujari, C., Aiswarya, Shetty, N.P.: Comparison of classification techniques for feature oriented sentiment analysis of product review data. In: Data Engineering and Intelligent Computing, pp 149–158 (2017)
15. Vladimir, N.V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
16. Naïve Bayes. https://cran.r-project.org/web/packages/naivebayes/naivebayes.pdf
17. David, M.: Support Vector Machines: The Interface to libsvm in package e1071 (2017)
18. Package Class. https://cran.r-project.org/web/packages/class/class.pdf