Measurement: Sensors 27 (2023) 100768

Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

Human behavior analysis on political retweets using machine learning algorithms

Het Patel *, Aditya Kansara, Boppuru Rudra Prathap, Kukatlapalli Pradeep Kumar

Computer Science and Engineering (Specialization in Data Science), Christ University, Bangalore, India

ARTICLE INFO

Keywords:
Political retweet
Opinion mining
Text classification
Twitter
Computing methodologies
Machine learning algorithms

ABSTRACT

The exponential rise in the use of social media has resulted in a massive increase in the volume of unstructured text created. This content is presented through messages, conversations, postings, and blogs. Microblogging has become a popular way for people to share what they are thinking. Many people express their thoughts on various issues relating to their hobbies. As a result, microblogging websites have become a valuable resource for opinion mining and sentiment research. Twitter is a well-known microblogging network, with over 500 million new tweets posted daily. The goal of this study was to mine tweets for political sentiments. The extraction of tweets relating to India's well-known political leaders of different states & parties in India and applying the polarity detection analysis of human behavior on the retweeted messages As a result, the sentiment classification algorithm is designed to determine whether tweets are more likely to predict the popularity of certain politicians among the general public. The subjectivity and polarity present in the tweets of political leaders are compared. The engagements of these leaders are then taken into account to determine their popularity. All these comparisons are then portrayed using data visualizations.

1. Introduction

People use social media as a platform to express their opinions, thoughts, sentiments, and emotions. Over the past few years, social media has been increasingly popular. Are you aware that on Twitter, more than 6000 tweets are posted every second? Because of the magnitude of the data from social media, it is quite difficult to comprehend the most recent trends and describe the general sentiments, polarity, and opinions of the people about any item and product. Twitter has become a vital communication tool for people from many different backgrounds thanks to its speed and ease of use. Social media users share their views, ideas, and opinions on a wide range of subjects. Twitter in India has become a platform where people conduct political discourse, considering that many politicians use Twitter to communicate with huge masses. Indian political leaders are among the most followed politicians around the world, and with this massive following, the corpus of text generated while replying to or retweeting the tweets of these leaders is also huge. Tweets from ten prominent Indian political leaders, as well as replies to those tweets, were collected and analysed to determine the general sentiment behind those tweets. We tried to keep the list of politicians diverse, so we decided to take several popular political parties and then select a leader from that party with the largest number of followers. We also tried to find out the polarity and subjectivity present in the tweets of those leaders. where polarity means the amount of positivity or negativity present in a particular tweet of a leader and subjectivity measures whether the tweet posted is factually correct or is a personal opinion of that leader. Then we compared the engagements of different political leaders and how they affected the image of a particular leader in the eyes of the public. This is done by measuring positivity among the replies made to particular tweets of a leader. In this paper, we investigate the function of sentiment analysis on political Twitter.

2. Literature review

Ansari et al. [1] explain that it is possible to evaluate the annotated corpus over itself using LSTM and various machine learning models, and findings using LSTM and Random Forests are encouraging. It is necessary to extract a larger number of tweets in order to improve this type of analysis and inference. In order to balance the distribution of classes among all the classes, the appropriate sampling techniques must be used. To create a robust corpus, semi-supervised corpus creation methods might be used. It was discovered that various TSA-related

* Corresponding author.
E-mail address: het.patel@btech.christuniversity.in (H. Patel).

https://doi.org/10.1016/j.measen.2023.100768
Received 30 September 2022; Received in revised form 21 January 2023; Accepted 2 May 2023
Available online 10 May 2023

works have used similarity metrics to find related news stories, themes, and user personalities by Adwan et al. [2] They also think that there is a good chance to integrate social network analysis and similarity metrics with current TSA methodologies, particularly machine learning approaches, as they are working with social networks. Wilson et al. [3] found out that part-of-speech features may not be helpful for sentiment analysis in the microblogging space, according to their experiments on sentiment analysis on Twitter. To ascertain if the POS features are just of poor quality as a result of the tagger's output or whether POS features are simply less effective for sentiment analysis in this domain, more research is required. Severyn et al. [4] described the deep learning approach to Twitter sentiment analysis on both the message and phrase levels. They provided a thorough explanation of their three-step procedure, which is the secret to their success, for training the network's parameters. Modern performance on both the phrase-level and message-level subtasks is shown by the resulting model. Their solution is the best on both subtasks when the average rank of all test sets, including progress test sets, is taken into account. Zhang et al. [5] came up with an augmented lexicon-based method specific to the Twitter data, which was then applied to perform sentiment analysis. Empirical experiments showed that the proposed method was highly effective and promising.

Twitter posts contain both rich context and event-based information that can be leveraged for the prediction of criminal incidents and activities. Prathap et al. [6] proposed a method on sentiment analysis on crime related tweets. Sentiment analysis determines the author's or reader's feelings about a text. Their stories affect politics and daily life. Prathap et al. [7] proposed a method for polarity detection using newspaper data as a source. Mix tweets in a multi-party setting to capture users' political feelings. AparupKhatu et al. [8] forecast users' political leanings from their tweets during the 2014 Indian General Election. 0.15 million people sent 2.4 million tweets. The relationship between mixed tweeting and political leanings is investigated. Neural network-based algorithms predict political leanings. Mix-tweeting reveals Twitter users' political views, according to our analysis. Politicians are using digital media to reach out to voters more and more. Twitter is a big and important part of political conversations. Twitter shaped [9] how people felt during India's 17th Parliament Elections. Corruption claims were a key strategy for political parties. In Ref. [10], Justin Paul et al. present a comprehensive survey of sentiment analysis techniques, issues, and trends. In the study, the methods and applications of sentiment analysis are examined. The pros and cons of the chosen methodologies are then contrasted and analysed. Future steps are outlined, and the issues with sentiment analysis are examined. Tu MyDoan et al. [11] propose defining political opinions, explaining why they're important, and discussing how well existing algorithms do it. Current approaches can't repair The author of Political Viewpoint Investigations (PVI) also examines PVI's challenges and makes research proposals. Matteo Cinelli et al. [12] proposed that users collaborate to address information cascades. Users who collaborate can distribute their messages more swiftly and efficiently. The length of the cascade is related to the number of people who initiate it. There comes a time when even the assistance of more experienced users is no longer useful. Automatically classify users into coordinated and uncoordinated groups.

3. Proposed system

We propose a model based on sentiment analysis for extracting the polarity present in political tweets made by popular Indian political leaders and the replies made to those tweets. Once the tweets and replies are scraped from Twitter, the tweets and their metadata are stored in data frames, while the replies made to those tweets are stored in a dictionary. The data is then cleaned to remove any noise present in the data. Then we used various natural language processing methods to perform analysis on the cleaned data to extract useful information. The information was then displayed using different data visualization methods. In order to do the experiment, we followed a 4-step procedure as follows: As shown in Fig. 1, data extraction is followed by data preprocessing, data analysis, and data visualization. The detailed procedure is explained in the following section.

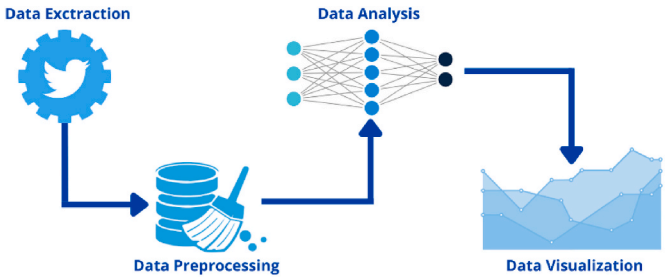

Fig. 1. System Architecture.

3.1. Data extraction

We used Twitter's official API to extract data from the platform. We had created the account and authorised an app from Twitter to collect the data. This application authenticates our account and issues a consumer key and access token, which are used to connect with the Twitter app. Once we have the access token, our app is free to interact with Twitter's API on behalf of the user. Then, using Twitter's API, we scraped the tweets of 10 Indian political leaders. We then stored the useful data, like user id, username, tweet id, tweet, language, replies count, likes count, retweets count, and quotes count, in a data frame. Data frames provide a readable format for the tweets and the related information.

To extract the replies made on particular tweets, we had to scrape them from Twitter's website. Scraping is an automatic method to obtain large amounts of data from websites. To implement it, we used Selenium, which is an open-source umbrella project for a range of tools and libraries aimed at supporting browser automation. Using Selenium, we scraped the replies made on the tweets that were present in the data frame using the tweet id. The scraped replies were then stored in the dictionary. Dictionaries are used to store data values in key-value pairs, which helped us map all the replies to their unique tweet ids.

3.2. Data preprocessing

Data preprocessing is a technique that is used to transform the raw data into a useful and efficient format to enhance the performance of the model in the analysis phase. The data we get from social media sites such as Twitter is highly unstructured, which makes preprocessing an essential and important step before the data is ready for analysis and generating some meaningful insights from the huge amount of information. The data that is scraped often contains typos, bad grammar, usage of slang, and the presence of unwanted content like URLs, stopwords, expressions, etc. So, it is very necessary to preprocess the text before working on it.

Here, preprocessing has been done in the following steps: Firstly, we will load the dataset and also import the necessary libraries. We then translate all tweets to English using the Google Translate API. The stop words that were used in the tweets must then be eliminated. Prepositions, pronouns, and other terms such as "being," "is," "the," "having," etc. are excluded. Since they are neither positive nor negative, they can be removed without altering the overall sentiment of the tweet. The next step involves removing from the tweet any unnecessary hashtags, URLs, and other symbols.

3.3. Data analysis

In this project, we use the Twitter-Roberta-base-sentiment-latest Transformer Model [13]. The transformer model is used to score the

Table 1
Final results of sentiment analysis.

Name	Tweet Length	Likes	Replies	Retweets	Quotes	Polarity	Subjectivity
Narendra Modi	198.528	25058	769	4400	127	0.154	0.306
Rahul Gandhi	155.152	45088	3357	10393	708	0.038	0.163
Arvind Kejriwal	174.23	6906	685	1034	118	0.089	0.263
Shashi Tharoor	197.523	1357	77	173	20	0.200	0.418
Basavaraj Bommai	217.256	791	25	59	3	0.042	0.079
Akhilesh Yadav	155.775	12373	828	2086	130	0.014	0.023
Asaduddin Owaisi	227.167	2711	85	641	21	0.029	0.297
Jagan Mohan Reddy	168.240	5851	309	916	47	0.191	0.323
Mamata Banerjee	215.869	4167	344	699	69	0.234	0.406
Yogi Adityanath	203.655	12275	452	1883	65	0.001	0.001

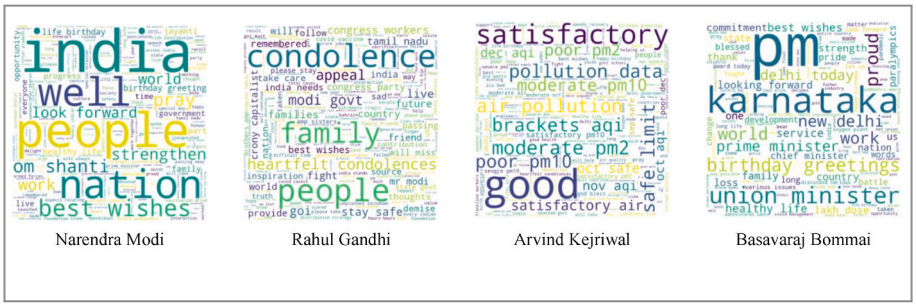

Fig. 2. Wordclouds of 4 leaders.

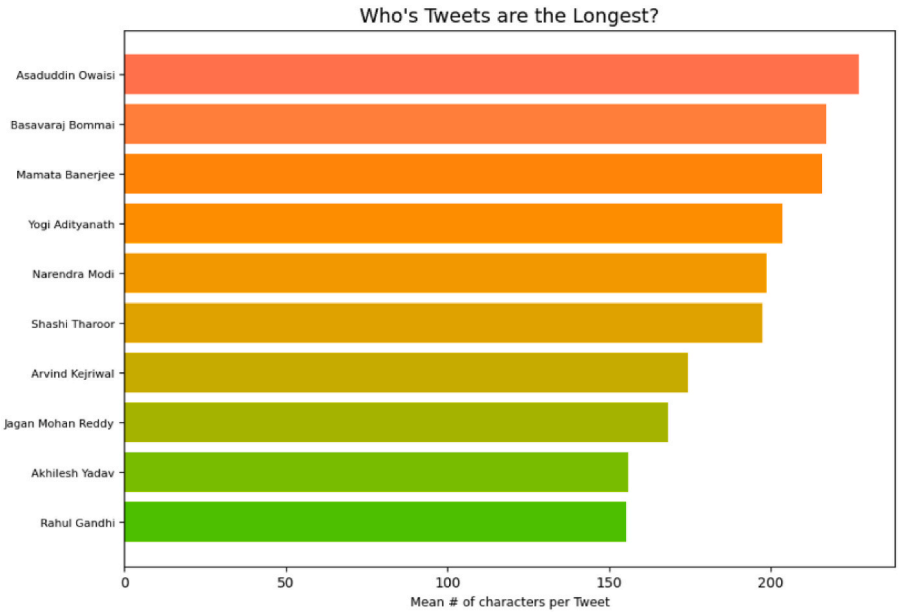

Fig. 3. Tweet lengths of different political leaders.

replies made on tweets based on the emotions of the user. The basic idea behind sentiment analysis is to determine whether a response is negative or positive and then compute the average response sentiment. Then polarity and subjectivity detection are performed on the tweets of politicians using Textblob. When a sentence is passed into Textblob, it gives two outputs, which are polarity and subjectivity. Polarity is the output that lies between [−1, 1], where -1 refers to negative sentiment and +1 refers to positive sentiment. Subjectivity is the output that lies within [0, 1] and refers to personal opinions and judgments. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information.

3.4. Data visualization

The result of the analysis is then shown visually in the form of bar charts, word clouds, scatter plots, and a sunburst chart. Effective visualization makes complex data more comprehensible and understandable by enabling people to view the outcome visually. The words that appear frequently are given more emphasis in word clouds. It aids in a more accurate assessment of a certain politician's viewpoint.

4. Results & discussion

Here we considered 10 influential Indian political leaders and

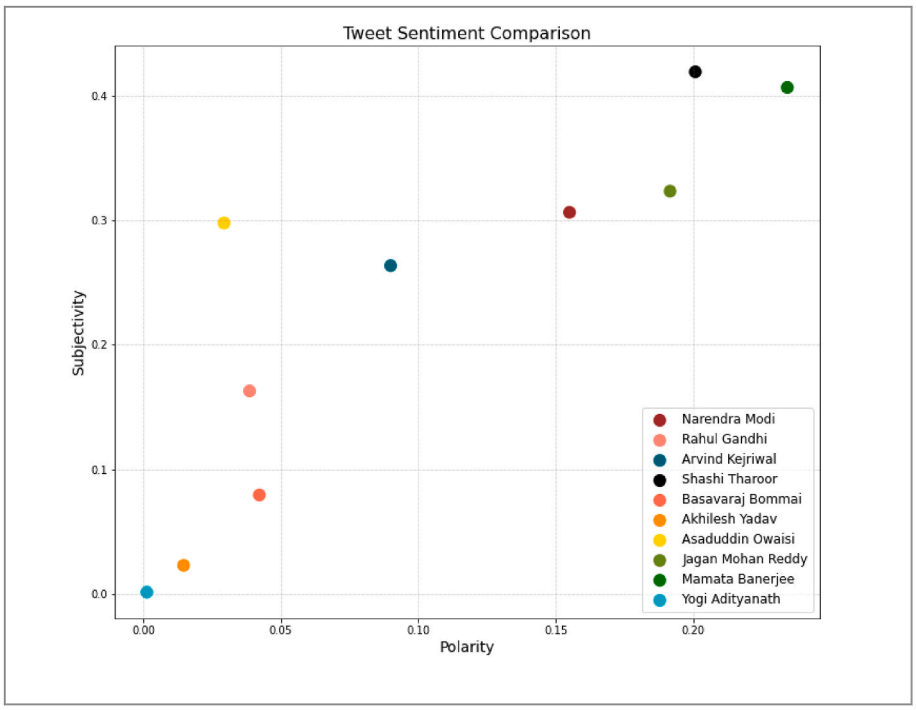

Fig. 4. Subjectivity and Polarity amongst the tweets of all 10 leaders.

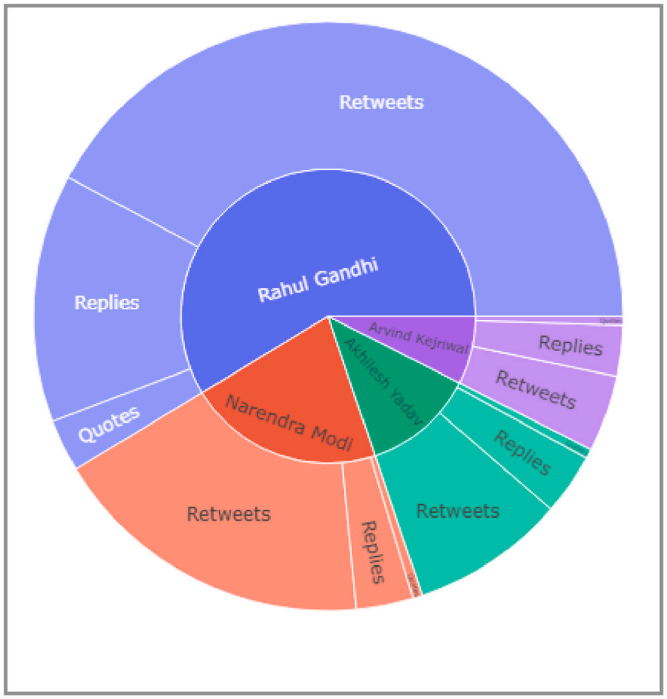

Fig. 5. Engagement of 4 leaders.

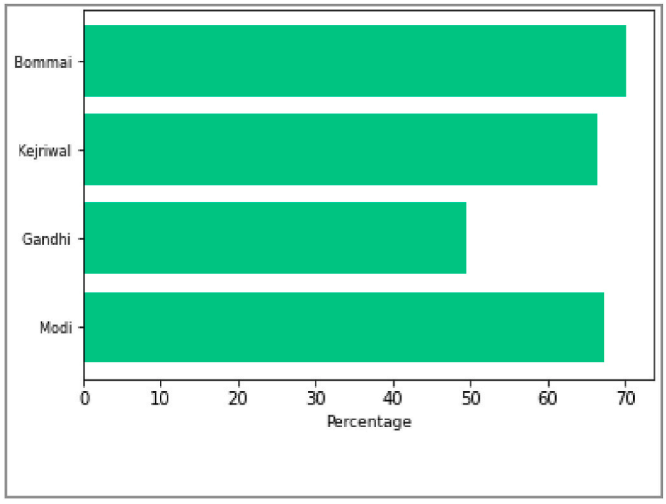

Fig. 6. Positivity in replies of 4 leaders.

performed sentiment analysis on their tweets and replies. The final result of the analysis can be seen in Table 1. Then we plotted the results into various graphs using data visualization.

From the above figures, we can infer several things, such as that from Fig. 2, we can find out which political leader posts the longest or shortest tweets, and the average tweet length is in the range of 150–225. Then, using word clouds, we found out the most common words occurring in the tweets of four different political leaders (Fig. 3). Next, we plotted the polarity and subjectivity of all leaders on a scatter plot (Fig. 4), and we found that the polarity of all leaders is above 0, which means that all leaders are posting positive tweets in general. And the subjectivity of all leaders is below 0.5, which means most of the leaders are posting factual statements rather than personal opinions. Fig. 5 depicts four leaders' engagements. Fig. 6 shows the positivity in the responses of four leaders. According to Figs. 5 and 6, the more engagement a leader has, the less positivity there is in the responses he or she receives.

5. Conclusion

Twitter sentiment analysis is a significant method to identify different user tweets. The suggested approach encourages examining and rating a range of political tweets and the comments made on those tweets based on various tweet properties. It is observed that all the leaders chosen generally post positive tweets, and most of those tweets are factually based and not subjective. It is also observed that the leader

with most engagement has the least amount of positivity in the replies made to his tweets. As demonstrated, transformer models can be used to analyse sentiment in a large corpus of tweets. Data visualization techniques were employed to comprehend the collected data. The results of these studies provide evidence for a relationship between political sentiment on Twitter.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] M.Z. Ansari, M.B. Aziz, M.O. Siddiqui, H. Mehra, K.P. Singh, Analysis of political sentiment orientations on twitter, Proc. Comput. Sci. 167 (2020) 1821–1828.
[2] O. Adwan, M. Al-Tawil, A. Huneiti, R. Shahin, A.A. Zayed, R. Al-Dibsi, Twitter sentiment analysis approaches: a survey, Int. J. Emerg. Technol. Learn. (iJET) 15 (15) (2020) 79–93.
[3] E. Kouloumpis, T. Wilson, J. Moore, Twitter sentiment analysis: the good the bad and the omg, 1, in: Proceedings of the International AAAI Conference on Web and Social Media, 5, 2011, pp. 538–541.
[4] A. Severyn, A. Moschitti, Unitn: training deep convolutional neural network for twitter sentiment classification, in: Proceedings of the 9th Inter*national Workshop on Semantic Evaluation*, 2015, June, pp. 464–469. SemEval 2015.
[5] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, B. Liu, Combining lexicon-based and learning-based methods for Twitter sentiment analysis, HP Lab., Tech. Rep.HPL-2011 89 (2011) 1–8.
[6] B.R. Prathap, K. Ramesha, Twitter sentiment for analysing different types of crimes, in: 2018 International Conference on Communication, Computing and Internet of Things (IC3IoT), 2018, February, pp. 483–488 (IEEE).
[7] D.J. Hemanth, Polarity detection on real-time news data using opinion mining, Intellig. Syst. Computer Tech. 37 (2020) 90.
[8] A. Khatua, A. Khatua, E. Cambria, Predicting political sentiments of voters from Twitter in multi-party contexts, Appl. Soft Comput. 97 (2020), 106743.
[9] J. Paul, N. Parameswar, M. Sindhani, S. Dhir, Use of microblogging platform for digital communication in politics, J. Bus. Res. 127 (2021) 322–331.
[10] M. Birjali, M. Kasri, A. Beni-Hssane, A comprehensive survey on sentiment analysis: approaches, challenges and trends, Knowl. Base Syst. 226 (2021), 107134.
[11] T.M. Doan, J.A. Gulla, A Survey on Political Viewpoints Identification, 30, Online Social Networks and Media, 2022, 100208.
[12] M. Cinelli, S. Cresci, W. Quattrociocchi, M. Tesconi, P. Zola, Coordinated inauthentic behavior and information spreading on twitter, Decis. Support Syst. (2022), 113819.
[13] D. Loureiro, F. Barbieri, L. Neves, L.E. Anke, J. Camacho-Collados, Timelms: Diachronic Language Models from Twitter, 2022 *arXiv preprint arXiv:2202.03829*.