A Bi-level Assessment of Twitter Data for Election Prediction: Delhi Assembly Elections 2020

Maneet Singh  
Indian Institute of Technology Ropar  
Rupnagar, India  
2018csz0008@iitrpr.ac.in

S.R.S. Iyengar  
Indian Institute of Technology Ropar  
Rupnagar, India  
sudarshan@iitrpr.ac.in

Akrati Saxena  
Department of Mathematics and Computer Science  
Eindhoven University of Technology  
Eindhoven, The Netherlands  
a.saxena@tue.nl

Rishemjit Kaur  
CSIR-Central Scientific Instruments Organization  
Chandigarh, India  
rishemjit.kaur@csio.res.in

ABSTRACT

Elections are the backbone of any democratic country, where voters elect the candidates as their representatives. The emergence of social networking sites has provided a platform for political parties and their candidates to connect with voters in order to spread their political ideas. Our study aims to use Twitter in assessing the outcome of the Delhi Assembly elections held in 2020, using a bi-level approach, i.e., concerning political parties and their candidates. We analyze the correlation of election results with the activities of different candidates and parties on Twitter, and the response of voters on them, especially the mentions and sentiment of voters towards a party over time. The Twitter profiles of the candidates are compared both at the party level as well as the candidate level to evaluate their association with the outcome of the election. We observe that the number of followers and the replies to candidates' tweets are good indicators for predicting actual election outcomes. However, we observe that the number of tweets mentioning a party and the temporal analysis of voters' sentiment towards the party shown in tweets are not aligned with the election result. Moreover, the variations in the activeness of candidates and political parties on Twitter with time also not very helpful in identifying the winner. Thus, merely using temporal data from Twitter is not sufficient to make accurate predictions, especially for countries like India.

CCS CONCEPTS

- Information systems → Social networks;  
- Networks → Online social networks.

KEYWORDS

Indian Election, Social Media, Temporal Analysis.
The second line of research in predicting the outcomes of an election aims at either analyzing the behavior of candidates on social media platforms or assessing the voters’ response towards the candidates on these platforms. One of the studies [23] was conducted on the 2010 USA midterm election using the tweets of more than 300 candidates as well as their followers. The engagement of candidates with the voters (or followers) on Twitter in the form of retweets and mentions was found to be linked with the outcome of an election [23]. Bright et al. [6] analyzed the use of Twitter by around 600 candidates of UK national elections held in 2015 and 2017. The vote percentage of candidates was found to be positively correlated with Twitter usage. According to Franzia [10], the surprising victory of Donald Trump during the 2016 presidential election could be associated with his supremacy in the use of free media. The tweets mentioning the candidate name can be used to identify the popularity of the candidate with the help of sentiment analysis as shown in the study analyzing the 2017 French elections [22]. The direct mentions of a candidate name may not be sufficient; instead, using additional mentions through different possible aliases of the names along with election relevant keywords may result in better predictions [11]. The network of users surrounding a candidate on social media through friends and followers can also be relevant for the election outcomes; however, it has very little evidence [7].

There is a lack of research that analyzes the election data at the bi-level, i.e., in terms of both participating political parties and the candidates. The bi-level approach is significant for elections where a large number of candidates are involved in any given election, and election outcome is dependent on the win or loss of those candidates; for example, the election process in India. Moreover, the need for undergoing a bi-level assessment can be understood based on the following two scenarios. First, a candidate could win because of their party or party leaders’ popularity and reach among the voters. Second, a candidate may win because of their persona among the voters. Our study uses the above approach on Twitter data to identify its role in predicting election winners. For our analysis, we collect and utilize a three-fold dataset related to the Delhi assembly elections held in 2020. Our dataset comprises generic tweets relevant to elections, candidate related tweets, and party related tweets. In our paper, we study the correlation of different parameters, such as tweet mentions, sentiments of voters, and availability and activity of different Twitter profiles, with the election results. We also study the changes in candidates’ and parties’ activities overtime before the election and users’ responses to them.

2 DATASET

We collect the Twitter data for the Delhi Assembly elections, 2020 using the Twitter API. We mainly collect three different kinds of datasets pertaining to Delhi Assembly elections as explained below.

2.1 Election relevant tweets

The tweets related to assembly elections held in Delhi, capital of India, containing keywords ‘Delhi’ and ‘Election’, were collected. Our collection process resulted in 162,556 tweets spanning from 22nd December 2019 to 7th February 2020, i.e., for a period of 48 days before the election date (8th February 2020). Later, we retained tweets of only those users that belong to Delhi, with the help of geo-location mentioned in their tweets or the city specified explicitly by them in the location field of their profile. The data from the location field has been found to be a valid source in past studies [12, 16]. There were finally 143,616 tweets in our dataset belonging to Delhi users.

2.2 Candidate related tweets

First, we manually collected the official Twitter handles of the candidates of three major political parties, i.e., BJP (Bhartiya Janta Party), AAP (Aam Aadmi Party), and Congress (Indian National Congress). Since all the candidates were not active on Twitter, we were able to extract 101 Twitter handles out of 203 candidates belonging to all three parties. These Twitter handles were then used to collect all the tweets by the candidates along with the retweets and replies to their tweets, from 21st January 2020 to 7th February 2020 using the streaming services by Twitter. The data collection was started from 21st Jan as all three parties declared their final list of candidates for the elections on this date [3]. On segregating the tweets related to candidates, there were a total of 14,323 tweets by the candidates themselves along with 928,545 retweets and 323,434 replies to their tweets.

2.3 Party related tweets

The Twitter handles of all three political parties specific to Delhi (@BJP4Delhi, @AAPDelhi, and @INCDelhi) were used to collect all their tweets as well as retweets and replies to their tweets using Twitter Streaming API. In this way, we were able to collect 4,733 tweets by all three parties along with 281,209 retweets and 40,893 replies to their tweets.

3 METHODOLOGY

3.1 Tweets Sentiment Extraction

All election relevant tweets containing only single party mentions were used for extracting sentiments with the help of the LIWC (Linguistic Inquiry and Word Count) tool [20]. LIWC tool uses a dictionary for extracting various linguistic features. Since we are interested in identifying sentiments in the tweets, we only extract word counts related to positive and negative emotions for each tweet. These word counts were normalized by dividing with the total word count of a tweet, given as one of the features in the tool. Using this, we were able to obtain the proportion of both positive as well as negative emotions in all the tweets.

3.2 Extracting Twitter Profile Attributes

The various attributes of candidates and political parties profiles were extracted from Twitter and are divided into two categories.

3.2.1 Long term attributes. The long-term attributes are not specific to the pre-election period but rather related to the overall profile on Twitter. These include account age, number of tweets, likes, followers, and friends on Twitter. All these attributes were collected using the Twitter REST API.

3.2.2 Short term attributes. The short-term attributes include number of tweets by an account, the average number of retweets and
4 RESULTS AND DISCUSSIONS

In this section, we will discuss our main findings. For better understanding of the analysis, it is vital to know the actual outcomes of the Delhi Assembly Elections held in 2020. The party that emerged as the winner was AAP (won 62 out of 70 seats), whereas BJP was the second best party (won 8 out of 70 seats), and the Congress did not win any seat in the elections.

4.1 Political Party Mentions

We first analyze the impact of the tweets mentioning a party on the election results.

4.1.1 Using Hashtags and Mentions. The use of hashtags or mentions are quite common on Twitter for escalating the outreach of any given tweet [9, 19]. Therefore, the top twenty hashtags and mentions, based on their usage frequency in our dataset are extracted and shown in Fig. 1. We observe that the top hashtags in the collected tweets contains all kind of hashtags, neutral as well as specific to any party or party members. The most frequently occurring hashtag specific to a party is DelhiWithBJP, which does not coincide with the actual results of the election. In the case of mentions, the two most frequently used Twitter handles were ArvindKejriwal and AamAadmiParty, and they both belong to the winning party.

4.1.2 Keywords based Analysis. Since mentioning a party need not necessarily be done through symbols '# or '@'; hence a keyword-based approach is used with the help of names and abbreviations of all three political parties. There were total 91,700 tweets that stated the name of any political party; hence a comparison of the number of tweets mentioning the parties is shown in Fig. 2(a). The party with the maximum number of tweets mentioning its name is BJP (Bharatiya Janta Party), whereas the party that won the election is AAP (Aam Aadmi Party).

4.1.3 Using Tweets Mentioning only one Party. As one tweet may refer to more than one party, therefore we extracted 62,978 tweets mentioning only single party [18] to get a more clear picture in terms of party mentions. The comparison between the three political parties is shown in Fig. 2(b). Here, it was again observed that BJP was getting more mentions than the winning party. However, this analysis does not consider the sentiments of the tweets to verify if the mentions are in the support of the party or not; we will discuss this later.

4.2 Analyzing Sentiments towards Party

The number of tweets mentioning any party just gives an indication of the most talked about party on Twitter and may not highlight the fact that whether the discussions were in favor of the party or not [18]. Hence, it is vital to study the sentiments towards those parties, as done in previous works [5, 18]. The tweets mentioning a single party were used for analyzing sentiments using lexicon based approach, i.e., explained in detail in Section 3.1. The mean
proportion of both positive and negative emotions are shown for all political parties in Fig. 3. The figure shows that both positive and negative sentiments are more for BJP as compared to other parties.

In the above analysis, the complete data of 48 days window was used for comparing sentiments towards each party. Assuming that the time difference from the election might affect the resultant sentiments, we perform a time-based analysis of sentiments towards political parties (Fig. 4). As seen from the figure, the proportion of positive sentiments is mainly dominated towards BJP for all different window sizes, except for the last three days, where Congress was leading in the positive sentiment. The winning party AAP does not lead in any window size for positive sentiments. In the case of negative sentiments, if we consider only the last one week, AAP had the lowest average proportion with respect to BJP (Effect-Size = 0.46 and p-value ≤ 0.0001) and Congress (Effect-Size = 0.47 and p-value ≤ 0.0001), where Effect size and p values are computed using Mann Whitney U Test [14]. Similarly, for the last two weeks, AAP again had the least proportion of negative sentiments as compared to both BJP (Effect-Size = 0.43 and p-value ≤ 0.0001) and Congress (Effect-Size = 0.48 and p-value ≤ 0.0001). These observations conflict with previous studies that mainly rely on sentiments (positives, negatives, and both) for predicting elections [5]. Our results show that the lower negative sentiments as compared to higher positive sentiments are a better indication of the winning.

4.3 Candidate’s Availability on Twitter
All candidates from three political parties were not available on Twitter; hence it becomes vital to compare the level of participation at the party level. A candidate is considered available if there is at least one tweet from the candidate during our collection period. The winning party was found to have maximum share of candidates available on Twitter (64%), followed by BJP(54%) and Congress(33%). This could be one of the reasons that might have made AAP candidates popular among the voters. Next, we compare the activeness of different participants on Twitter. In our analysis, the activeness refers to the number of tweets by the candidates during our data collection period. The activeness is studied from two perspectives, first on a daily basis and second using multiple time frames, each spanning over three days.
Figure 6: Activeness of candidates of different political parties on each 3 days window spanned over 18 days of data collection.

Figure 7: Comparing Twitter Profiles of the winning and loosing candidates.

Figure 8: Comparing Twitter Profiles of candidates party-wise.

Figure 8: Comparing Twitter Profiles of candidates party-wise.

Hence, we used a window of three days as one time frame to average out the frequent changes in the number of candidates tweeting (at least once), from each political party, on the platform daily. The time frame comparison of percentage of candidates’ participation on Twitter is shown in Fig. 6. As seen, the percentage of candidates active on Twitter does not vary much for all the different time frames. There was slight variation for the AAP party; it gets less involved in tweeting as we move closer to the election. On the other hand, Congress was relatively less involved during the mid-interval and this might be due to the major political campaigns in that period by other parties i.e AAP and BJP [2].

4.4 Comparing Twitter Profiles

Although an election might be conducted on a single day or a span of few days or even in months, a candidate or a party may run their political campaign for a longer period to widen their reach among voters. The comparison of various attributes of Twitter profile (Refer Section 3.2 for identifying profile details) is performed from the following two perspectives.

4.4.1 Candidates. The Twitter profile of the winning candidates is compared with the losing candidates using the mean of all the attributes, as shown in Fig. 7. The average number of followers is higher for the candidates who won the election than those who lost the election.
were scaled for displaying on one plot, where the scaling factor for each attribute is defined below the axis label itself. The use of online social networks by political parties for propagating their agendas and ideologies is getting very common. In this study, we use a bi-level approach to analyze the correlation of Twitter activities with the election winners both at the party level and candidate level. We observed that the widely used measures such as party mentions and sentiment towards the party were unable to identify the winning party. In the case of Twitter profiles of the candidates, the winning candidate has relatively more followers and a higher number of the average replies per tweet. In this work, we have used the Twitter data for one election; however, in the future, researchers can assess the capabilities of Twitter by performing a cross-national study, covering countries with different economic and demographic backgrounds.

REFERENCES

[22] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics 1947, 50–60.