HASHTAGS ARE (NOT) JUDGEMENTAL: THE UNTOLD STORY OF LOK SABHA ELECTIONS 2019

Saurabh Gupta, Asmit Kumar Singh, Arun Balaji Buduru and Ponnurangam Kumaraguru

Indraprastha Institute of Information Technology - Delhi (IIIT-Delhi)
{saurabhg,asmit18025,arunb,pk}@iiitd.ac.in

ABSTRACT

Hashtags in online social media have become a way for users to build communities around topics, promote opinions, and categorize messages. In the political context, hashtags on Twitter are used by users to campaign for their parties, spread news, or to get followers and get a general idea by following a discussion built around a hashtag. In the past, researchers have studied certain types and specific properties of hashtags by utilizing a lot of data collected around hashtags. In this paper, we perform a large-scale empirical analysis of elections using only the hashtags shared on Twitter during the 2019 Lok Sabha elections in India. We study the trends and events unfolded on the ground, the latent topics to uncover representative hashtags and semantic similarity to discover sentiments during elections. We collect over 24 million hashtags to perform extensive experiments to find the trending hashtags, and cross-reference them with the tweets in our data set to list down notable events. We also use semantic similarity based techniques to find related hashtags and latent topics among the hashtags.

Index Terms— Social networking sites, Social tagging systems, Hashtags, Big Data Analysis

1. INTRODUCTION

Online social media platforms like Twitter are being used by people to spread information and opinions among other users. A lot of times, people are observed reporting the ground events happening near them, making Twitter a source of getting breaking news [1, 2]. For example, when the terrorist attacks in Mumbai in 2008 were happening, Twitter users in India (especially in Mumbai) were providing an instant eyewitness account of what was happening at the ground [3]. More recently, a lot of media channels covered the reactions of people all over India using Twitter when article 370 was scraped [4, 5]. Twitter is considered so effective that even the Indian government recently asked them to remove accounts spreading rumors about Kashmir [6].

More recently, hashtags brought together the users who are concerned about Lok Sabha elections and wanted to share their opinions. Twitter has become one of the most effective unofficial platforms to share news, opinions, facts, fake news, and a political playground with #LokSabhaElections2019 among the top three most tweeted hashtags in 2019 [7]. Hashtags also aided researchers to study such political events from multiple perspectives like participation in #iranelection [8], retweet behavior on real-world ground events [9], temporal and demographic characterstics [10], category and nature of users and tweets [11] and so on. To study political hashtags, [11] collected data based on a single hashtag #cdnpoli and analyzed content related to Canadian politics on Twitter.

In this paper, we perform a large-scale empirical study of political hashtags from Indian context on Twitter. We start with finding the most trending hashtags over the course of elections and during each phase. We then map these trends to real-world events that happened and were captured in our dataset. Further, we use semantic similarity to find related hashtags, and perform LDA on the complete dataset to find out topics during the elections. During 2019 Lok Sabha elections in India, Twitter was used by a lot of political parties, candidates, party supporters and common people to spread opinions, promotions, campaign, etc. We collected data from Feb 05, 2019 to Jun 25, 2019. Our collection process was heavily based on hashtags. We looked at hourly trends in keywords and hashtags to manually filter only the ones that are related to elections.

We believe our study can help several entities involved in political movements. The trends across all phases during elections can help political parties assess user sentiments towards them over Twitter and help to plan political propaganda. The patterns also facilitate the users to get a slight intuition about what party or candidate is more favored. The events fetched using hashtags give an idea about what is going on around that hashtag. The topics and semantics are majorly dominated by the candidates who heavily use social media to express their opinions. Semantic similarity also reveals the #hashtags to which a candidate's or party's name is associated. For example, we (in a completely unsupervised manner) observe that #modi is getting associated with #surgicalstrikes. Most of the analysis mentioned above is imaginable when you have a lot of attributes from the tweets. The fact that we only use the

bare minimum hashtags for all this makes this study different from others. We present a way to achieve similar results using just the hashtags.

2. DATA COLLECTION AND INITIAL ANALYSIS

In this section, we first discuss the data collection strategy, then showcase some preliminary analysis on hashtags.

2.1. Data Collection

The Loksabha Elections in India started on Apr 11th, 2019 and ended on May 19th, 2019.[1] We collected tweets from Feb 05, 2019 to Jun 25, 2019 - based on the intuition that people start talking about elections way before the actual dates and go on talking about it several days after it gets over. The elections occurred in seven phases where votes were cast in a single day followed by a few no-voting days. The timeline is shown in Table 1.

Phase	Date of voting	Duration of each phase
1	Apr 11	Apr 11 - Apr 17
2	Apr 18	Apr 18 - Apr 23
3	Apr 24	Apr 24 - Apr 28
4	Apr 29	Apr 29 - May 5
5	May 6	May 6 - May 11
6	May 12	May 12 - May 18
7	May 19	May 19 - May 22[2]

Table 1. Phase-wise election's date and duration.

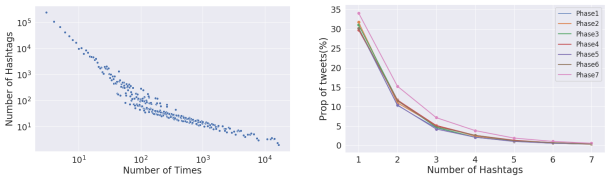
Initially, we started looking at hourly trends in keywords and hashtags from twenty-two cities in India. We manually selected hashtags related to elections based on the hourly trends and used Twitter's streaming API to get the following posts containing such hashtags. We collected a total of 45.1 million tweets, out of which 9.4 million were original tweets, and the rest were retweeted or quoted tweets. On investigation, we found some discrepancies with the collection process. For example, Twitter API failed to parse the hashtags from the text of some tweets. To resolve these discrepancies we: i) parsed the hashtags in instances of tweets where a user inserted a space between the # and the term. For example, # elections2019 is parsed as #elections, and ii) filtered the instances of tweets where the Twitter API failed to capture the hashtags. The parsing and filtering on the 9.4 million tweets resulted in the removal of 1.18 million tweets. In the remaining 8.22 million tweets, there were 24.9 million hashtags. Some statistics about the final 8.22 million tweets are given in Table 2. The data is publicly available at: `http://precog.iiitd.edu.in/requester.php?dataset=hashtag20`.

Total tweets after preprocessing	8,228,932
Total hashtags in tweets	24,958,397
No. manually curated unique hashtags	1,500
No. unique hashtags in dataset	970,408
Minimum no. of hashtags in a tweet	1
Maximum no. of hashtags in a tweet	7
Average no. of hashtags per tweet	3.02

Table 2. Summary Statistics of the dataset. Number of hashtags vary from 1 to 7 with an average of 3.02 hashtags per tweet.

2.2. Data Distribution

Figure 1(a) shows the distribution of the number of times hashtags are shared. The distribution follows power law, i.e., most hashtags are shared only a few number of times.



(a) Hashtag Frequencies (b) Phase-wise Distribution

Fig. 1. (Left) Both x and y axes are in log scale. There are more number of hashtags that are tweeted lesser number of times, and vice-versa. (Right) There are 40-45% of tweets that contain only one or two hashtags. The number decreases with the increase in the number of hashtags in each tweet.

For many experiments, we have divided the data into seven parts representing the seven phases. Each part has tweets corresponding to a particular phase duration. Figure 1(b) shows the distribution of number of hashtags in each tweet. Around 45% tweets contain one or two hashtags. There are comparatively less number of tweets that contain at least 5 hashtags.

3. TRENDS AND EVENTS

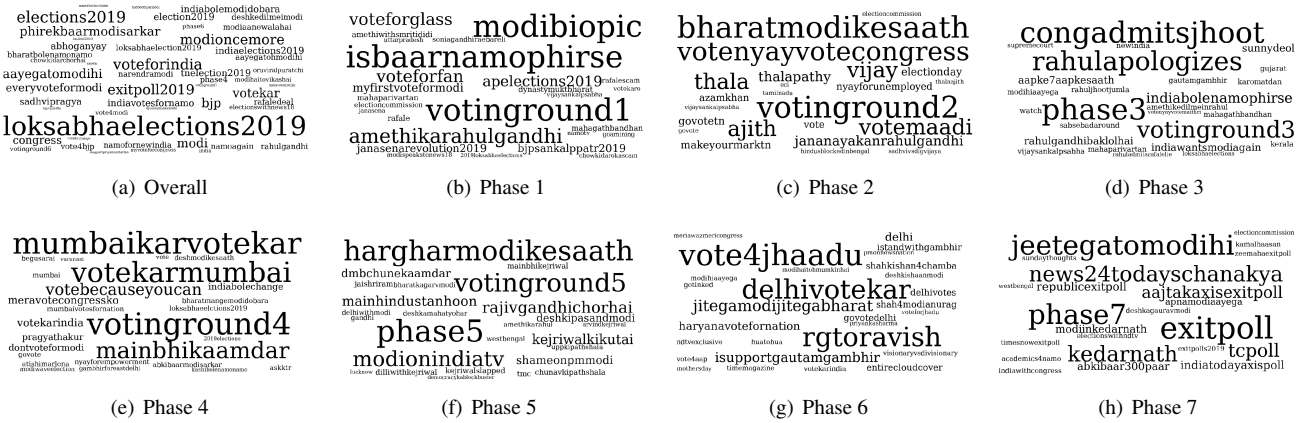
In this section, we first portray some general trends using word clouds. We then utilize some top trends to fetch tweets that reveals the event unfolded on the ground around those trends.

3.1. Trends

We draw a word cloud of the top 50 most occurring hashtags to show general trends in hashtags throughout elections.

[1] Except for the Vellore Parliamentary constituency in Tamil Nadu where the Election Commission of India (ECI) canceled the elections [12].

[2] The counting of votes started on May 23. Therefore, we assumed that Phase 7 lasted until May 22.



(a) Overall (b) Phase 1 (c) Phase 2 (d) Phase 3

(e) Phase 4 (f) Phase 5 (g) Phase 6 (h) Phase 7

Fig. 2. Top Hashtags over the course of elections and during each phases.

Figure 2(a) shows curiosity for elections among Twitter users as they were using hashtags like #votekar, #voteforindia, #loksabhaelections2019.

Next, we divided the dataset into seven parts depicting each phase to observe general trends in the duration of each phase. Each part contains all the tweets posted within the duration of that particular phase. Figures 2(b) - 2(h) shows trends during phases 1-7, respectively. We removed the hashtags that are common between the overall trends (Figure 2(a)) and phase-wise trends (Figure 2(b) - Figure 2(h)). The removal is necessary to differentiate overall trends from phase-wise trends. A behavior similar to the overall trends, is observed in the phase-wise plots as well.

Qualitatively, the sentiment around trending hashtags is mixed. On the one hand, we have hashtags like #modihaitomumkinhai, #abkibaarmodisarkar, #deshkipasandmodi, #apnamodiaayega, while on the other hand, there are trending hashtags like #saynotobjp, #shameonpmmodi, #indiawithcongress, etc.

3.2. Events

We wanted to find the notable events that happened during the course of elections using just hashtags. For this, we used the trending hashtags generated from the previous section to cross-reference it with our dataset in order to track actual events that happened on the ground. We make a list of hashtags that are shared more than 8,000 times and randomly choose them to track down events. We search our dataset using the chosen hashtags to make a list of tweets containing each hashtag. Then we randomly choose a tweet that contains at least two sentences. Please have a look at supplementary materials to find a few examples[3].

[3]Supplementary Material

4. TOPICS

In this section, we used approaches to find the topics hidden among hashtags during the course of elections. For this experiment, we filtered the tweets that have at least five hashtags. The number of hashtags in tweets vary from a minimum of 1 to a maximum of 7. After filtering, we were left with 212,935 tweets containing at least five hashtags. We treated each hashtag as a word and group of all hashtags in a tweet as a sentence. Further, we applied Latent Dirichlet allocation (LDA) to find latent topics.

We wanted to experiment and find what topics are prominently related to Elections on Twitter, in order to analyze the overall user sentiment towards a party, and understand the problems and opinions floating among Tweets. We used Latent Dirichlet Allocation (LDA) to find latent topics among the hashtags in an unsupervised manner. We chose LDA for the task because it calculates the document-topic and word-topic distributions using Dirichlet priors, which leads to better generalization [13].

Preprocessing To keep topic modeling based only on hashtags, we extracted the hashtags from each tweet. Each hashtag is treated as a token, i.e., a word. These tokens are then combined to form a sentence, which is a document.These documents are used to perform LDA. We remove the less frequent words and the stop words. The remaining words are reduced to their lemmas to make topics more meaningful [14]. The way we create a document conforms with the preprocessing steps necessary to get good topics using LDA.

Choosing no. of topics We need to choose the number of topics (K), which indicates how many latent topics the LDA model should find.[15] proposed a framework to calculate coherence measures to evaluate topic models. Therefore, to find the optimal number of topics to look for, we build many LDA models with a different number of topics ranging from 5 to 40 and calculated their coherence values. Same keywords start to repeat in multiple topics if the value of K is too large. There-

fore, we choose the value of K that lies at the starting point of the convergence of coherence values. The optimal value of K for our corpus is 20.

Topics After preprocessing the tweets and finding the optimal values for the number of topics, we run the LDA model on our dataset. We manually choose a qualitative keyword to represent the grouped #hashtags and assigned it as the topic. Following are some example of topics we were able to find:

- **Elections**: #elections2019, #loksabhaelections2019, #vote, #loksabhaelections
- **Promotions**: #voteforindia, #votekar, #vote4bjp, #vote4modi
- **Modi Praise**: #phirekbaarmodisarkar, #modioncemore, #modihaitomumkinhai, #namoagain

5. RELATED HASHTAGS

We wanted to find hashtags in our dataset that are semantically related to a query hashtag over the course of elections. We use the skip-gram model with negative sampling [16, 17] for the purpose. Skip-gram is used to maximize the similarity between the words which appear next to each other in the given corpus. It creates a continuous vector for each word in a manner that preserves a word's context.

We used cosine distance as the similarity metric and chose four hashtags: #modi, #raga, #bjp, #congress to perform a qualitative analysis to find hashtags that are most semantically similar to these four. We chose these hashtags because the Bhartiya Janata Party (BJP) and the Indian National Congress (or just congress) are two biggest parties with 435 and 420 candidates [18] participating in the elections, respectively. The hashtag #modi refers to Narendra Modi (BJP leader), and #raga refers to Rahul Gandhi (Congress leader). Table **??** shows the results for #modi, #raga, #bjp, #congress. As we can see, the hashtags found from the experiment indeed are quite similar, e.g., #raga is similar to #pappu [19], #robertvadra [20] and #bjp is similar #modisarkar, #sambitpatra (BJP candidate), and #rss [21].

6. WHO'S WINNING THE SOCIAL MEDIA BATTLE?

We wanted to quantify the reach of a candidate on Twitter based on hashtags. Note that, with the metric $influence$ proposed in this section later, we are not predicting the outcome of an election. It is used just to understand the sentiment going around about a candidate. We first searched for "vs" and "versus" in the hashtags present in our dataset to find the ones that represent two candidates competing against each other in elections. Our exhaustive search ended up getting us 8 candidates, but we looked at the top three most prominent ones like #smritiiranivsrahulgandhi, #gautamgambhirvsatishimarlena, and #sadhvipragyavsdigvijay.

If $hashtags[1, 2, .., n]$ represents a list of semantically similar hashtags, we find $C_{hashtag[i]}$, the number of times a hashtag i had appeared. To add sentiment information, we manually annotated the semantically similar hashtags as positive, negative or neutral. Then we find the number of occurrences, $C^{+}_{hashtag[i]}$ for all positive and $C^{-}_{hashtag[i]}$ for all negative hashtags for all candidates to calculate the influence, given as:

$$influence = \sum_{i=1}^{x} C^{+}_{hashtag[i]} - \sum_{i=1}^{y} C^{-}_{hashtag[i]}$$

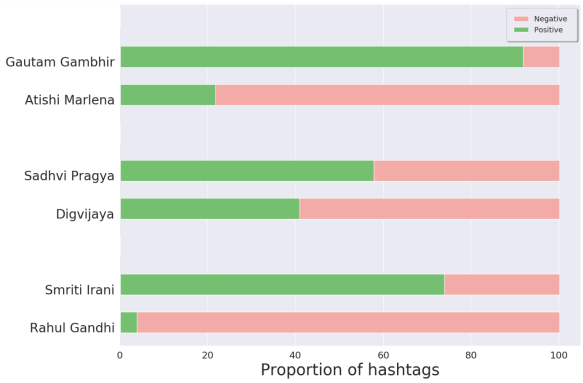


Fig. 3. The percentage of negative and positive hashtags for key battles. The proportion of positive hashtags for each winning candidate is greater than the negative one, rendering them a higher influence score.

Figure 3 shows that the losing candidates have more negative hashtags than the positives ones. In other words, the influence score of loosing candidate is negative. The same pattern is true for the other two candidates as well.

7. DISCUSSION

In this paper, we perform a large-scale empirical study of political hashtags from Indian context on Twitter. We wanted to analyze what all patterns we can observe using only the hashtags from data collected during Loksabha elections 2019 in India. We collected 24.9 million hashtags from 8.22 million tweets for this study. We first show the trending hashtags from each phase. We take one of the most trending hashtags in a phase and cross-reference to our dataset to find the actual event that unfolded on the ground. Moreover, we use Latent Dirichlet Allocation (LDA) to assign qualitative topics among hashtags. The topics help us qualitatively cluster the hashtags. We then use the skip-gram word embeddings to find that #pappu is semantically similar to #raga, and #modi is accredited with #airstrike, #surgicalstrike.

8. REFERENCES

[1] Beatriz Arias, "How newsrooms can use twitter's latest tools to break news.," .

[2] Abhinav Sharma, "Why twitter is still the best place for breaking news despite its many challenges.," .

[3] Claudine Beaumont, "Mumbai attacks: Twitter and flickr used to break news," .

[4] DH Web Desk, "Twitter reacts to govt's decision on art 370," .

[5] PTI New Delhi, "Centre revokes article 370 : Virtual sloganeering takes over twitter," .

[6] Rezwan, "Indian government asks twitter to remove accounts spreading rumours about kashmir," .

[7] Nandita Mathur, "Twitter celebrates 12th birthday of the hashtag," .

[8] Devin Gaffney, "Iranelection: Quantifying online activism," in *In Proceedings of the Web Science Conference (WebSci10*, 2010.

[9] Meenakshi Nagarajan, Hemant Purohit, and Amit Sheth, "A qualitative examination of topical tweet and retweet practices," 2010.

[10] Munmun De Choudhury, Shagun Jhaver, Benjamin Sugar, and Ingmar Weber, "Social media participation in an activist movement for racial equality," *Proceedings of the ... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media*, vol. 2016, pp. 92–101, 05 2016.

[11] Tamara A. Small, "What the hashtag?: A content analysis of Canadian politics on Twitter," *Information Communication and Society*, vol. 14, no. 6, pp. 872–895, 2011.

[12] Manasa Rao, "As vellore lok sabha election looms closer, aiadmk desperate for a win," .

[13] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, 2003.

[14] C. Jacobi, W.H. van Atteveldt, and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," *Digital Journalism*, vol. 4, no. 1, pp. 89–106, 2016.

[15] Michael Röder, Andreas Both, and Alexander Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, New York, NY, USA, 2015, WSDM '15, pp. 399–408, ACM.

[16] Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean, "Efficient estimation of word representations in vector space," 01 2013, pp. 1–12.

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, USA, 2013, NIPS'13, pp. 3111–3119, Curran Associates Inc.

[18] Manish Kanadje and Ankita Nanda, "Analysis of the contesting candidates in general election 2019," .

[19] Prabhash K Dutta, "How rahul gandhi became 'pappu' of politics," .

[20] India Today Web Desk, "Private citizen robert vadra appears on hoardings featuring sonia gandhi outside congress headquarters," .

[21] Admin, "Relation between rss and bjp," .