#IVoted to #IGotPwned: Studying Voter Privacy Leaks in Indian Lok Sabha Elections on Twitter

No Author Given

No Institute Given

Abstract. Online Social Networks (OSNs) play a crucial role in elections worldwide. Users post their opinions and sentiments on events, candidates, and parties. One of the cardinal principles in elections is to ensure that the party (or candidate) to which a citizen vote remains secret. However, given that citizens are free to express their opinions and views on OSN platforms like Twitter, some of them, in direct and indirect ways, reveal their political inclinations, which we refer to as Voter Privacy Leaks (VPL). In this paper, we cross-link VPL user's online details with other publicly available information (like electoral rolls, which is a list of all eligible voters), to get access to their personally identifiable information (PII) including their voter ID, age, gender, address, and family details. Finally, to safeguard such users, we develop browser plugin based nudge that leverages machine learning-based classifiers to flag a given post on Twitter as a VPL or a Non-VPL, thereby helping users protect their voter privacy. To validate our approach, we focus on the period of Lok Sabha elections held in 2019 in India, the largest democracy in the world. We collect tweets starting from April 11, 2019, to May 22, 2019. We detect 91,253 instances of VPL and using a subset of electoral rolls, successfully cross-link 44 Twitter users to their exact PII. Our proposed nudge detects 93% of VPL incidents.

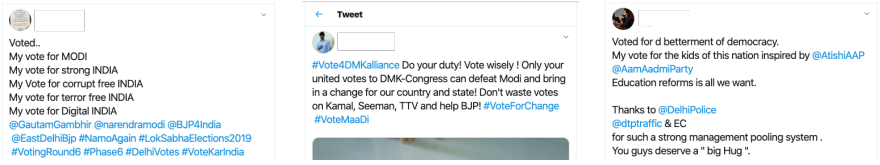
Keywords: Deanonymization · Privacy Flaw · Twitter

1 Introduction

As a microblogging service, Twitter is 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 [30]. 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 [2]. The platform is considered so pivotal that even the Indian government recently asked Twitter to remove accounts spreading rumors about Kashmir [26]. More recently, Twitter brought together the users who are concerned about the Lok Sabha elections 2019 in India and wanted to share news, opinions, facts, fake news,

and became a political playground with #LokSabhaElections2019 among the top three most tweeted hashtags in 2019 [19].

Voter Privacy, as given in Section 39 in The Conduct of Elections Rules, 1961, is defined as the maintenance of secrecy of voting by electors within polling station and voting procedure [7]. Voter privacy is implemented by a method called the secret ballot. The secret ballot is a voting method in which a voter's choices in an election or a referendum are anonymous, forestalling attempts to influence the voter by intimidation, blackmailing, and potential vote buying [27]. The origin of the secret ballot in Indian elections is traced back to section 39 in the conduct of elections rules, introduced in 1961. The acts state that "Every elector to whom a ballot paper has been issued under rule 38 or any other provision of these rules shall maintain the secrecy of voting within the polling station". The voters cast their vote in isolation and exercise their right to voter privacy.



(a) Example reveal of BJP (b) Example reveal of INC (c) Example reveal of AAP

Fig. 1. Users reveal their votes while posting tweets on Twitter. We have taken examples of the top three most notable political parties in India, viz., BJP, INC, and AAP.

Twitter users lose their voter privacy by posting tweets that reveal the name of the party or the candidate they support. In this work, we study such tweets, and refer to them as **Voter Privacy Leaks (VPL)**. Figure 1 shows examples of tweets where users have posted VPLs during Lok Sabha Elections in 2019. A user revealing their vote warrants their eligibility to vote, and consequently, the presence of their personal information in electoral rolls. Electoral rolls are a publicly available list of all the voters curated in PDF files and categorized by their states and constituencies[1] maintained by the Election Commission of India (ECI). A single file contains the address of a part that belongs to a constituency and a complete list of voters in that part. Also, it provides several personally identifiable information like voter ID, age, gender, address, and family details along with their names. Identification of an individual using cross-linking the two data sources poses a severe threat like identity theft, blackmail, reputation damage, unwanted disclosure and regret, and so on [16, 32, 10, 37, 3]. The individuals become prone to targeting for several political gains [12, 33] or just for threatening them based on whom they support. Considering the increasing intol-

[1] https://eci.gov.in/electoral-roll/link-to-pdf-e-roll/

erance and lynch mobs in India, such tweets successfully linked with personally identifiable information might even lead to a life and death situation [8, 14].

Previous research on studying leaks on Twitter primarily explores tweets: related to everyday phenomena like vacations, drinking, and diseases [17]; exposes violation of privacy settings [20]; build tools to detect personally identifiable information [6, 5, 15], and so on. Unlike the general causes, breaching user privacy to target them for political gains is commonplace as several cases have unfolded in recent years [12, 21]. Therefore, in this paper, we cross-link Twitter details to electoral rolls to reveal several personally identifiable information. Then we detect and prevent users from posting tweets (VPLs) that might compromise their privacy using a browser-based visual nudge. Our experiments are centered around the following research questions:

- Users on Twitter, directly or indirectly, post VPLs. The Election Commission of India (ECI) releases electoral rolls that contain several personally identifiable information (PII) like age, gender, family details, etc. Cross-linking the two can pose serious privacy threats, which takes us to the first research question we ask:
  RQ1. [**Cross-Linking**] *Using a Twitter profile and a selection of tweets, can we successfully find PII of an individual in the electoral rolls?*
- When Twitter users are linked with electoral rolls, their single tweet becomes a threat to their privacy. They become vulnerable to identity theft, unwanted disclosure, discrimination, and so on [32]. Therefore, for awareness and mitigation, we ask the second research question:
  RQ2. [**Detection and Protection**] *How can we detect a voter privacy leak from a tweet to inform the user and prevent it from happening in the future?*

We successfully cross-link Twitter data set with a subset of electoral rolls to find 44 exact matches, which is a big enough number to prove VPLs to be a severe privacy threat. We discuss the cross-linking methodology with a case study of a Twitter user in Section 3. To mitigate the issue, we formulate a binary classification problem with two classes: "VPL" and "Non-VPL" tweets. A random forest classifier with count vectors performs the best in all evaluation metrics. Using the trained classifier model, we develop a browser extension that nudges users whenever they are posting a tweet that might make them lose their voter privacy.

Privacy and Ethics We hope that the Election Commission of India appreciates the research and stop making the electoral rolls publicly accessible. We understand that there are serious privacy concerns while conducting the experiments, and therefore, we have taken several measures to keep them in check. We discuss the detailed ethical considerations in Section 5.

2 Background and Dataset

In this section, we give a background about the Lok Sabha Elections 2019 proceedings, candidate registration, and describe the data set of tweets we collected using Twitter for this work.

2.1 Lok Sabha Elections 2019

The Lok Sabha in India is the lower house of India's bicameral parliament. It has 545 members elected through direct election. To participate in elections, candidates from several political parties fill and submit an affidavit to the Election Commission of India (ECI). The affidavit contains information about the candidate's background, tax information, assets, social media handles, and so on. For this work, we were interested in the candidate's name, constituency (location), and their social media handles(specifically Twitter username) to collect tweets that mention either of these things. The Bhartiya Janata Party (BJP) and the Indian National Congress (INC) are the two major national parties in India, with 435 and 420 candidates participating in the Lok Sabha elections in 2019.

The Lok Sabha Elections in India started on Apr 11th, 2019. They ended on May 19th, 2019[2]. The elections occurred in seven phases, where votes were cast in a single day, followed by a few no-voting days. We collected tweets using Twitter APIs from the start of phase 1 of elections, i.e., April 11, 2019, until the counting started. The counting of votes started on May 23. Therefore, we assumed that the last phase, i.e., phase 7, lasted until May 22. The timeline is shown in Table 1.

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

Phase	**Date of voting**	**Duration of each phase** [3]
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] Except for the Vellore Parliamentary constituency in Tamil Nadu, where the Election Commission of India (ECI) canceled the elections [25].

[3] We make the split based on an assumption that when users vote in a phase, they will talk about things related to corresponding phase unless a new phase starts. For example, during phase 1, the trending hashtags will be related to phase 1 unless phase 2 starts.

2.2 Dataset

Twitter Usernames of Candidates We extracted candidate's information from their corresponding affidavits available at ECI's website [4]. The affidavits were available in PDF format, and we manually downloaded all of them to list down the candidate names, their constituencies (for location), and their Twitter display name and username. We also listed down Twitter handles of BJP and INC for the data collection process.

Voter Privacy Leaks (VPLs) To collect relevant tweets to answer our research questions, we used the Twitter APIs with the following heuristic:

- We continuously monitored Twitter during the elections and focused on tweets in which users have declared that they voted. We followed several trends to collect tweets relevant for this work:
  1. The popular/trending hashtags started by media channels, like #KBPMselfie[5], which asks people to post their inked finger selfies after voting, thus implying that they voted.
  2. We manually discovered hashtags that indicate that the person voted. For example, #gotinked, #ivoted, #firstvote, and so on[6].
- We filtered tweets in which a user has mentioned any of the following entities:
  1. the name of a candidate they favor.
  2. the Twitter handle of a candidate they favor.
  3. the name of a party they support.
  4. the Twitter handle of a party they support.
  5. a hashtag with any of the four above, for example, #myfirstvoteformodi, #votedforcongress, etc.

We used the filter to segregate promotional tweets from the actual users revealing their votes. The filter helped us collect tweets relevant to answer the research questions we proposed.

Non-VPLs We also collected an equal number of tweets in which a user has revealed neither the candidate nor the party they support. We randomly sampled these tweets over the course of data collection to get more general representation of Non-VPL data to be used to answer our research questions. We kept the numbers equal to eliminate the class imbalance problem.

3 Cross-Linking

For our investigation to find a matching entry in electoral rolls to reveal PII, starting from Twitter details, we collect and parse entries from electoral rolls. In this section, we explain the data collection process and cross-link them to answer our first research question.

[4] https://affidavit.eci.gov.in/
[5] https://twitter.com/abpnews/status/454135087194337280
[6] Extensive list of hashtags to be added in supplementary upon acceptance

3.1 Background

For every constituency, there is a list of voters known as the electoral roll. An electoral roll of a constituency includes information of all the registered voters of the particular constituency to which it belongs. The available information includes several personally identifiable attributes such as name, age, gender, father's/husband's name[7], address, and the Electors Photo Identity Card (EPIC) number, commonly known as voter ID. The Election Commission of India (ECI) makes electoral rolls publicly accessible for all the 543 constituencies in India. Electoral rolls allow responsible authorities to: i) verify the details of a voter, ii) streamline the voting process on the day of polling, and iii) ensure that a person cannot vote twice. Due to the availability of so many attributes that can uniquely identify an individual, publicly available electoral rolls pose a threat to an individual's privacy. To protect individuals from the threat, countries like Australia do not allow online browsing of these electoral rolls [23].

3.2 Case Study

Dataset India has 543 Lok Sabha constituencies divided among its 29 states and 7 union territories. The Election Commission of India (ECI) provides state-wise downloadable electoral rolls in PDF format[8]. For our analysis, we used electoral rolls from the National Capital Territory (NCT) of Delhi[9], which includes 7 Lok Sabha constituencies where each constituency has 10 assembly constituencies, summing a total of 70 constituencies. The 70 constituencies are further divided into several parts based on the region and population. The electoral rolls are available for each part of the 70 constituencies in PDF format. We downloaded these electoral rolls by manually finding patterns in the URL structure. Figure 2 shows the pattern we sequentially followed to automate the download of electoral rolls. There is a captcha available on the website, but it does not protect the URLs.

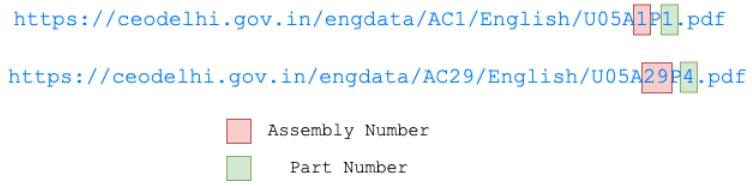


Fig. 2. The URL pattern used to download electoral rolls. We sequentially changed the assembly numbers and part numbers to get all the data.

[7] Husband's name in case of married females, otherwise father's name
[8] https://eci.gov.in/electoral-roll/link-to-pdf-e-roll/
[9] Delhi is officially called the National Capital Territory of Delhi. For official events like Elections, the neighboring cities are also included as part of Delhi.

We downloaded the electoral rolls for a total of 4,504 parts of 22 assembly constituencies of New Delhi. Since the entries were in PDF format, we performed optical character recognition (OCR) on them using Tesseract [31]. Tesseract is a widely used and freely available OCR engine released under the Apache License[10]. We have used several regular expressions on Tesseract output to map information about several users in a PDF correctly. From all the downloaded PDFs, we obtained the details of around 4.4 million voters. Some statistics about our data set is given in Table 2.

We only use a subset of all the collected data to prove our claim, but it can easily be extrapolated to the complete data and for details on more users as well.

Table 2. Summary statistics of the electoral rolls data set.

Total voters whose information is extracted	4,437,907
Assembly Constituency with most voters	Vikaspuri
Assembly Constituency with least voters	Mundka
Average no. of voters per constituency	201,723
Attributes about an individual extracted from the PDFs	name, age, gender, father's/husband's name, address, and voter ID

Cross Linking to predict Identity We have started from a tweet where users have mentioned that they voted for a candidate or a party. The fact that they voted ensures their eligibility to vote, and therefore, their information must be present in the electoral rolls. We cross-link the Twitter display name with the names present in the electoral rolls.

For the case study, we limited the experiments to only the tweets that either mention NCT of Delhi as a location or any of the candidates participating in the Lok Sabha elections from some part of NCT of Delhi. The filtering is done to avoid downloading personally identifiable information about a lot of users. We filter out the tweets that were posted in phase 6 on 12th of May 2019, i.e., the day of polling in NCT of Delhi (see Table 1) assuming that majority amount of leaks happen on the day of voting itself. After applying the filters mentioned above, we got 2,130 tweets from the complete dataset. However, the only personally identifiable information present in the twitter profile of a user is their Twitter display name and their location. It was challenging to find an exact match in the corresponding electoral rolls data set. Therefore, we only keep the tweets where the user has a unique Twitter display name. Out of 2,130 tweets, we find 1,041 tweets where the user has a unique name. Similarly, in the case of the electoral rolls data set, out of 4.4 million entries, there are 783,550 unique names.

Using the 1,041 tweets and 783,550 unique data points, we use the following heuristics to cross-link the tweets to electoral rolls data set:

[10] https://github.com/tesseract-ocr/tesseract

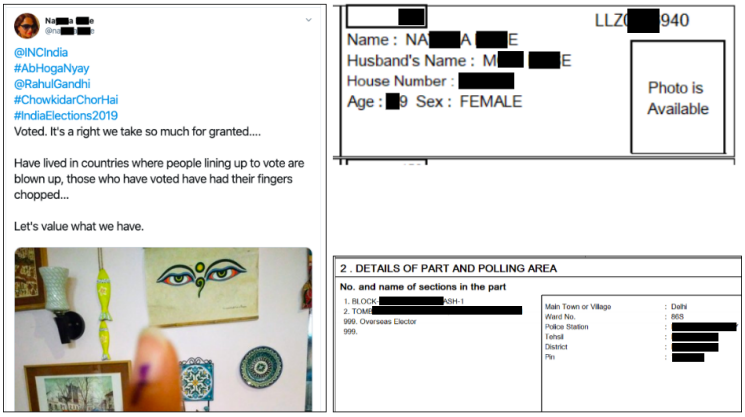


Fig. 3. An example of cross linking experiment. A Twitter user NaXXXa XXXe tweeted about their vote, hence, revealing their preferences towards a party and losing their voting privacy. The Twitter display name is successfully linked to their entry in the electoral rolls. We have censored their voter ID, names, and house number from the results.

1. The name in electoral roll and name in twitter profile (after prepossessing[11]) should be same.
2. The revealed entity (candidate's name/twitter username) in the tweet should be related to the constituency whose electoral roll is being considered.
3. The father's/husband's name is present in the following/followers of the user who posted the tweet.

After using the heuristics for cross-linking as mentioned above, we derived 44 exact matches, i.e., the user who posted the tweet was successfully mapped to their entry in electoral rolls data set, hence, giving us access to their personally identifiable information, which ideally should be private and non-accessible. Figure 3 shows an example of such a tweet made by NaXXXa XXXe (Twitter username: @naXXXaXXXe). Note that we have randomly chosen an entry from the 44 matches to show our results. We successfully linked their name with the details in the electoral rolls. The house number added to the details of part and polling area allows to locate their exact address of residence. Using the EPIC No. (or voter ID) alone, i.e., LLZXXX940, one can easily browse these details online at https://electoralsearch.in. Note that some details are redacted to protect their privacy and avoid information leakage.

The implications of cross-linking to reveal personal information, especially the address of a user, can lead to severe consequences. In the past, disclosure of location through Twitter led to a burglary at a user's house [11]. Other studies suggest that such disclosure of personal information may cause identity theft,

[11] We used normalization techniques like converting to lowercase, removing special characters, emojis, etc. from the names.

blackmail, reputation damage, unwanted disclosure and regret, and so on [16, 32, 10, 37, 3]. The opaqueness of passive data collection on Twitter-like platforms can lead to severe consequences. For example, researchers have characterized users who tweet under the influence of alcohol (and users revealing their medical conditions). They argued that such information shared with law enforcement agencies (and medical agencies) might get you arrested (and allow them to increase your medical bills) [13, 17].

4 Detection and Prevention

The importance of detecting Voter Privacy Leaks (VPLs) on Twitter and preventing them from happening cannot be emphasized enough. Therefore, in this section, we discuss our methodology to answer the second research question. More specifically, we use machine learning algorithms to detect VPLs and build a browser extension that helps prevent users from them.

4.1 Classification

When most users tweet sensitive information, they are unaware of the negative repercussions that it might have [28]. As a step towards educating users and shielding them against VPLs, we train a binary classification model that identifies if a given tweet has any content that can be used to reveal PII about the user if cross-linked with electoral rolls. Further, we use the model and build a browser extension for users to provide a real-time nudge to users while they draft their tweet. The nudge is in the form of a visual warning that is displayed based on the content of the tweet.

Dataset All the 91,253 tweets we collected in Section 2.2, where a user revealed the party or the candidate they support along with a voting hashtag are labeled as "VPL" tweets. We also randomly sampled the same number of tweets that **do not** contain any information that reveals the user's preferences over all phases of the election using similar Twitter APIs. We labeled them as "Non-VPL" tweets. Combining the two types of tweets gave us a total data set of 182,506 tweets with an equal number of "VPL" and "Non-VPL" tweets. We consider an equal number of both types of tweets to avoid the problems that arise due to class imbalance. The data helps us form a binary classification problem with two classes - "VPL" and "Non-VPL".

Classifier We performed pre-processing on the text before generating their word embedding. First, we sanitized the text in the tweets by removing URLs and emojis. Then we also remove the occurrences of special characters like # and @. On Twitter, hashtags always are prefixed by the # symbol and mentions by the @ symbol. Therefore, the words that follow these characters have high

significance because the data is collected using them. Finally, we randomly split the tweets into training and testing sets of 127,754 and 54,752, respectively.

We used the Scikit Learn module for building our classifiers [24]. We used two feature selection techniques: Count Vectorizer and Term Frequency - Inverse Document Frequency (TF-IDF). The Count Vectorizer builds a vocabulary of all the words in the tweets and then creates vectors for tweets where each index of the vector stores the count of occurrences of a word in that tweet. On the other hand, TF-IDF is an extension of the Count Vectorizer as it normalizes the count of each word in the tweet by taking into account the count of the words across all tweets. We used two types of classification models: Naive Bayes Classifier and Random Forests [4]. Naive Bayes is often considered an excellent baseline for text classification tasks [36]. We train these models with the classes being *VPL* and *Non-VPL*. Table 3 shows the resulting accuracy, precision, recall, and F-measure of all four variant models. The best performing model in terms of all metrics is a Random Forest Classifier with Count Vectorization as a feature selection technique. Count Vectorization is better because a tweet is a micro blog, and after all the preprocessing, frequent words like party/candidate names in hashtags/mentions get more weightage.

Table 3. Performance of various classification methods, using different learning algorithms and embedding.

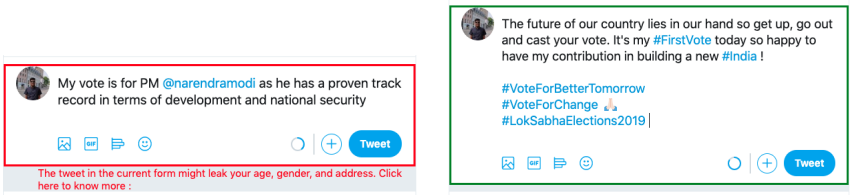
Classification Model	**Accuracy**	**Precision**	**Recall**	**F - measure**
Naive Bayes + TFIDF	0.80	0.82	0.80	0.80
Naive Bayes + Count Vectors	0.81	0.83	0.81	0.81
Random Forest + TFIDF	0.90	0.91	0.91	0.91
Random Forest + Count Vectors	**0.93**	**0.94**	**0.94**	**0.94**

4.2 Nudge

Preventing users from posting "VPL" tweets require building solutions that are useful and gets user attention. Inspired from the works that created nudges for other platforms like Facebook [37], we have used design choices that are proven to be effective [29, 1]. We designed a modification to Twitter's web app that shows visual warnings to nudge users to consider the content of their tweet and make it a "Non-VPL" tweet. We build a browser extension to implement the nudge. While a user drafts their tweet, the extension runs the classifier model on the tweet text in real time and nudges the user against providing any sensitive information. The nudge in the form of a visual warning is color coded as:

– Red: if the classifier predicts the current text as the "VPL" class.
– Green: if the classifier predicts the current text as the "Non-VPL" class.

Figure 4(a) shows the case where the current text is classified as a VPL. The text box becomes red and a warning is displayed to prompt the user. Figure 4(b) show the other case where the current text is classified as a Non-VPL. The text box remains green in this case, unless user enters something that might be revealing. The example text used in both figures is taken randomly from the test set.



(a) VPL (b) Non-VPL

Fig. 4. (Left)The extension creates a red background and shows a nudge message below while drafting the tweet if the classifier reports a leak. In this case, the user has revealed that his vote is for a specific candidate. (Right) The extension creates a green background in this case. This is because the tweet doesn't reveal any sensitive information and is safe to be posted.

5 Ethical Considerations

We have used two data sets in this work: i) Lok Sabha elections 2019 data from Twitter, and ii) Electoral rolls from Election Commission of India's (ECI) website. The data from Twitter was collected using the publicly available APIs, and we strictly followed the Twitter policies during data collection. The electoral rolls data from ECI's website was scraped. The use of data in this research warranted a justification of the ethical implications and the decisions made during the work. The methods followed during the study are influenced by the works around the internet research ethics and, more specifically, Twitter research ethics [38, 35, 9]. We handled data and made decisions in data collection in an ethical manner with no intent to harm any individual's privacy in any way [34, 22, 18]. Note that while reporting, we hide PII, and though the practice limits reproducibility, we choose not to release the data publicly. Further, we discuss the following aspects relating to this data: i) ethics in data collection, ii) risk-benefit trade-off in conducting this work.

5.1 Ethics in Data Collection

We have used Twitter's public APIs to collect data during the elections based on several rules (as discussed in 2.2). We did not scrape any data from Twitter during the collection process. We provided a user agent string that makes our intentions clear and provides a way for ECI web admins to contact us with questions or concerns. We requested data at a reasonable rate and strived never to be confused for a DDoS attack. We have saved only the data we needed from the page. We have respected the content and chose not to make it public in any condition. We ensured that we would respond in a timely fashion to any outreach we receive from ECI's website. The electoral rolls were stored on an encrypted hard disk accessible only to the researchers involved in the study.

5.2 Risk-Benefit Trade-off

We want to make users aware of their right to voter privacy with the primary objective of minimizing any potential harm. Secrecy in voting is essential to understand and maintain as it protects users from being targeted and radicalized. Our research provides an extensive observation of the consequences of publicly posting the vote in form of a VPL and provides a way to detect and prevent it if a user is vulnerable.

The most significant risk of this research is the loss of individuals' privacy. We have used multiple mechanisms to mitigate privacy concerns by using encrypted storage and making the decision not to share the data. The ECI website protects the privacy of website visitors, but due to unawareness, they are revealing a lot of information about voters in India. The most significant aim of this study is to make them aware of the consequences that can happen because they have made electoral rolls publicly accessible, and if possible, suggest them to use measures to protect electoral rolls as it is done in other countries like Australia [23].

6 Conclusion

In this work, we investigate the privacy concerns that arise when an individual loses their voter privacy by openly broadcasting their votes. Due to considerable interactions about Lok Sabha elections over Twitter, we specifically focus on tweets having special patterns, hashtags, or mentions, which when posted by a user, poses a threat to their privacy. We cross-link users from collected tweets to another data set of electoral rolls made publicly accessible by the Election Commission of India (ECI). The electoral rolls contain several PII like name, age, address, family details, and voter ID. The revelation can harm users in multiple ways like blackmail, identity theft, reputation damage, and so on. Therefore, we propose ways to mitigate these privacy concerns by: i) building a classifier model to detect tweets that can harm a user's privacy, and ii) developing a browser extension that works with Twitter's web app and shows a visual nudge to prevent user from posting anything that might make them lose their voter privacy.

References

1. Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L.F., Komanduri, S., Leon, P.G., Sadeh, N., Schaub, F., Sleeper, M., et al.: Nudges for privacy and security: Understanding and assisting users' choices online. ACM Computing Surveys (CSUR) **50**(3), 44 (2017)
2. Beaumont, C.: Mumbai attacks: Twitter and flickr used to break news (2008), https://www.telegraph.co.uk/news/worldnews/asia/india/3530640/Mumbai-attacks-Twitter-and-Flickr-used-to-break-news-Bombay-India.html
3. Boyd, D.M., Ellison, N.B.: Social network sites: Definition, history, and scholarship. Journal of computer-mediated Communication **13**(1), 210–230 (2007)
4. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324, https://doi.org/10.1023/A:1010933404324
5. Caliskan Islam, A., Walsh, J., Greenstadt, R.: Privacy detective: Detecting private information and collective privacy behavior in a large social network. In: Proceedings of the 13th Workshop on Privacy in the Electronic Society. pp. 35–46. ACM (2014)
6. Cappellari, P., Chun, S., Perelman, M.: A tool for automatic assessment and awareness of privacy disclosure. In: Proceedings of the 18th Annual International Conference on Digital Government Research. pp. 586–587. ACM (2017)
7. Dube, M., Jain, K.: Elections, law and procedures. Vedpal Law House (1985)
8. Dutta, P.K.: 16 lynchings in 2 months. is social media the new serial killer? (2018), https://www.indiatoday.in/india/story/16-lynchings-in-2-months-is-social-media-the-new-serial-killer-1275182-2018-07-02
9. Fiesler, C., Proferes, N.: "participant" perceptions of twitter research ethics. Social Media+ Society **4**(1), 2056305118763366 (2018)
10. Gross, R., Acquisti, A.: Information revelation and privacy in online social networks. In: Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society. p. 71–80. WPES '05, Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10.1145/1102199.1102214, https://doi.org/10.1145/1102199.1102214
11. Grove, J.V.: Twitter your way to getting robbed (2009), https://mashable.com/2009/06/01/twitter-related-burglary/
12. Hern, A.: Cambridge analytica: how did it turn clicks into votes? (2018), https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie
13. Humphreys, L., Gill, P., Krishnamurthy, B.: How much is too much? privacy issues on twitter. In: Conference of International Communication Association, Singapore. Citeseer (2010)
14. Khanna, P., Ghadyalpatil, A., Das, S.: Death by social media (2018), https://www.livemint.com/Politics/jkSPTSf6IJZ5vGC1CFVyzI/Death-by-Social-Media.html
15. Liang, H., Shen, F., Fu, K.w.: Privacy protection and self-disclosure across societies: A study of global twitter users. new media & society **19**(9), 1476–1497 (2017)
16. Lyon, D.: Surveillance society: Monitoring everyday life. McGraw-Hill Education (UK) (2001)
17. Mao, H., Shuai, X., Kapadia, A.: Loose tweets: an analysis of privacy leaks on twitter. In: Proceedings of the 10th annual ACM workshop on Privacy in the electronic society. pp. 1–12. ACM (2011)

18. Markham, A., Buchanan, E.: Ethical decision-making and internet research: Version 2.0. recommendations from the aoir ethics working committee. Available online: aoir. org/reports/ethics2. pdf (2012)
19. Mathur, N.: Twitter celebrates 12th birthday of the hashtag (2019), https://www.livemint.com/companies/news/twitter-celebrates-12th-birthday-of-the-hashtag-1566546599327.html
20. Meeder, B., Tam, J., Kelley, P.G., Cranor, L.F.: Rt@ iwantprivacy: Widespread violation of privacy settings in the twitter social network. In: Proceedings of the Web. vol. 2, pp. 1–2 (2010)
21. Michael Barton, P.D., McIntyre, N.: Digital election: what demographics are the parties targeting? (2018), https://www.theguardian.com/politics/2019/dec/03/digital-election-what-demographics-are-the-parties-targeting
22. Mislove, A., Wilson, C.: A practitioner's guide to ethical web data collection. In: The Oxford Handbook of Networked Communication. Oxford University Press (2018)
23. Orr, G., Mercurio, B., Williams, G.: Australian electoral law: a stocktake. Election Law Journal **2**(3), 383–402 (2003)
24. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
25. Rao, M.: As vellore lok sabha election looms closer, aiadmk desperate for a win (2019), https://www.thenewsminute.com/article/vellore-lok-sabha-election-looms-closer-aiadmk-desperate-win-106523
26. Rezwan: Indian government asks twitter to remove accounts spreading rumours about kashmir (2019), https://globalvoices.org/2019/08/16/indian-government-asks-twitter-to-remove-accounts-spreading-rumours-about-kashmir/
27. Rusk, J.G.: The effect of the australian ballot reform on split ticket voting: 1876–1908. American Political Science Review **64**(4), 1220–1238 (1970)
28. Sarikakis, K., Winter, L.: Social media users' legal consciousness about privacy. Social Media + Society **3**(1), 2056305117695325 (2017). https://doi.org/10.1177/2056305117695325, https://doi.org/10.1177/2056305117695325
29. Schaub, F., Balebako, R., Durity, A.L., Cranor, L.F.: A design space for effective privacy notices. In: Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015). pp. 1–17 (2015)
30. Sharma, A.: Why twitter is still the best place for breaking news despite its many challenges. (2019), https://www.forbes.com/sites/quora/2017/01/10/why-twitter-is-still-the-best-place-for-breaking-news-despite-its-many-challenges/
31. Smith, R.: An overview of the tesseract ocr engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). vol. 2, pp. 629–633. IEEE (2007)
32. Solove, D.J.: Understanding privacy, vol. 173. Harvard university press Cambridge, MA (2008)
33. Stokes, S.C.: Political clientelism. In: The Oxford handbook of political science. Oxford University Press (2007)
34. Townsend, L., Wallace, C.: Social media research: A guide to ethics. Aberdeen: University of Aberdeen (2016)

35. Vitak, J., Shilton, K., Ashktorab, Z.: Beyond the belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. pp. 941–953. ACM (2016)
36. Wang, S., Manning, C.D.: Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2. p. 90–94. ACL '12, Association for Computational Linguistics, USA (2012)
37. Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P.G., Cranor, L.F.: I regretted the minute i pressed share: A qualitative study of regrets on facebook. In: Proceedings of the seventh symposium on usable privacy and security. p. 10. ACM (2011)
38. Zimmer, M., Kinder-Kurlanda, K.: Internet research ethics for the social age: New challenges, cases, and contexts. Peter Lang International Academic Publishers (2017)