# NivaDuck - A Scalable Pipeline to Build a Database of Political Twitter Handles for India and the United States

Anmol Panda
anmol.panda777@gmail.com
Microsoft Research India
Bengaluru, Karnataka, India

A'ndre Gonawela
andregon@umich.edu
Gerald R. Ford School of Public Policy,
University of Michigan
Ann Arbor, Michigan

Sreangsu Acharyya
Microsoft Research India
srach@microsoft.com

Dibyendu Mishra
dibyendumishra96@gmail.com
Microsoft Research India

Mugdha Mohapatra
mumoha@microsoft.com
Microsoft Research India

Ramgopal Chandrasekaran
ramgopal@umich.edu
University of Michigan
Ann Arbor, Michigan

Joyojeet Pal
Microsoft Research India
joyojeet.pal@microsoft.com

## ABSTRACT

We present a scalable methodology to identify Twitter handles of politicians in a given region and test our framework in the context of Indian and US politics. The main contribution of our work is the list of the curated Twitter handles of 18500 Indian and 8000 US politicians. Our work leveraged machine learning-based classification and human verification to build a data set of Indian politicians on Twitter. We built NivaDuck, a highly precise, two-staged classification pipeline that leverages Twitter description text and tweet content to identify politicians. For India, we tested NivaDuck's recall using Twitter handles of the members of the Indian parliament while for the US we used state and local level politicians in California state and San Diego county respectively. We found that while NivaDuck has lower recall scores, it produces large, diverse sets of politicians with precision exceeding 90 percent for the US dataset. We discuss the need for an ML-based, scalable method to compile such a dataset and its myriad use cases for the research community and its wide-ranging utilities for research in political communication on social media.

## CCS CONCEPTS

• **Information systems** → **Social networking sites**; • **Human-centered computing** → *Empirical studies in collaborative and social computing*.

## KEYWORDS

twitter, politics, archive, india, united states


*SMSociety '20, July 22–24, 2020, Toronto, ON, Canada*

ACM ISBN 978-1-4503-7688-4/20/07…$15.00
https://doi.org/10.1145/3400806.3400830

**ACM Reference Format:**
Anmol Panda, A'ndre Gonawela, Sreangsu Acharyya, Dibyendu Mishra, Mugdha Mohapatra, Ramgopal Chandrasekaran, and Joyojeet Pal. 2020. NivaDuck - A Scalable Pipeline to Build a Database of Political Twitter Handles for India and the United States. In *International Conference on Social Media and Society (SMSociety '20), July 22–24, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3400806.3400830

## 1 INTRODUCTION

The election of Barack Obama to the presidency of the United States of America in the 2008 elections was a watershed moment for the use of new media in political campaigning [24, 39, 45, 48, 55, 82]. It established the potential to use new web-based media platforms for grassroots political organisation [50] and voter mobilization [17]. In their initial phase, social media platforms like Facebook and Twitter were credited with lowering the bar of participation for hitherto underrepresented communities [19, 85]. They allowed for more informal language in socio-political discussions [85], enabled collaborative participation by the electorate in political events [19, 85] and allowed them to shape the discourse as it evolved. On the other hand, they afforded politicians a relatively unregulated [64, 65] means to directly reach voters [64], without the mediation of mainstream electronic media organisations [19]. This has paved the way for their misuse by some parties and individuals to further their political agendas.

In recent years, polarization [40] and bias have become central to discussions around social media platforms [5]. These platforms have also been associated with condoning the spread of uncivil and extreme speech [59, 62], encouraging the dissemination of disinformation [1, 84], allowing trolling [71, 80, 91] and the use of bots [87] that impact both the opinion surrounding campaigns [71] and potentially their end outcomes [27].

Outside the US, social media has been seen to enable certain forms of spontaneous organizing that have contributed to resistance against states, as seen in the Arab Spring [2, 47]. In democracies, ideology-driven parties have used social media effectively to coalesce volunteers around certain ideas [11], target opponents,

and expand their political influence, as observed in Germany [4], Austria [90], Italy [12], France [6] and India [57].

In India, during the run-up to the 2014 general elections, then-PM-candidate Narendra Modi used platforms like Twitter and Instagram effectively, bypassing mainstream media to re-brand himself [64] from a right-wing strongman to a technology and internet-savvy modernist politician [63]. The landslide victory [79] of his Hindu-nationalist Bharatiya Janata Party (BJP) established [64] social media as a potent medium of political communication and branding in India. In the years since, politicians of all parties and hues have taken to various social media to engage with voters, promote their political agenda and critique their opponents. Moreover, the availability of cheap smartphones [88] and the rapid decline in cost of data / internet services [89] has allowed hundreds of millions of Indians to access social media and engage in political discourse. While the democratizing potential of these platforms is indeed laudable, the interplay between their affordances [59] and the extant socio-political environment in India [3] has allowed the viral dissemination of rumour [3] and extreme speech [85], often causing violence with casualties [3, 30].

The range of outcomes of social media interventions by bad actors, ranging from election manipulation in the US [25, 33, 91] to lynchings in India [3, 8, 18, 30, 58] has pushed researchers to closely consider the role of politicians on social media. These vary from understanding the content of politicians' online speech [36], the language and tone of their speech [69], the issues they choose to address (or ignore) [37] and the affordances of social media they engage in during major political events [52].

Early work on political social media looked at its use by election campaigns in western democracies with early adoption of social media use like the US (2008 and 2016) [20], Finland (2011) [81], Germany [44], Austria [21], Italy [13, 15, 86] and France [15], among others. Recent research has addressed some of the more existential crises related to social schism in Western states including issues around the use of social media to spread disinformation [25, 33, 91], false news [33], racist and anti-immigrant propaganda [22] and distrust in democratic institutions [71].

However, a large scale study of political actors beyond elected national figures has not yet been attempted, especially at the regional and grassroots levels in nations of the Global South. At a time when democracies face the dual challenge of rising populism in politics [23, 28] and distrust in governing institutions [46] spreading at scale [56, 78] on social media, we provide a means to fill this void through NivaDuck.

We proceed on the premise that a large scale database of political actors, particularly those in the party system, is important to understand how a social media political campaign operates at various levels. This is critical for diverse electoral systems, like that of India and the United States, in which a significant number of individual and collective political actors influence the discourse in different ways. Such systems also do not have easily collated lists of political actors. Aggregators like election commissions may not do a good job of keeping track of elected representatives' social media, and unelected politicians, particularly those who are not candidates for any public office but nonetheless devote themselves to active political work through party affiliations would not be tracked in any such public source.

We propose a method that combines machine-learning and human verification to build this database of politicians. We defined a politician in India as follows:

- Members of parliament (Lok Sabha and Rajya Sabha [1]), state legislatures, local governing bodies like Municipal Corporations and *grampanchayats* (village councils), etc.
- Unelected members of political parties such as party president, vice presidents, general secretaries, spokespersons, and anyone standing for elections on a party ticket
- Members of parties media, social media and It teams (colloquially referred to as IT wing or Social media cells)
- Official party handles at the national, state and district levels
- Members of student bodies or youth wings of political parties like the ABVP, NSUI, IYC, BJYM,etc.
- Grassroots party workers like booth agents, volunteers who publicly report themselves as members or workers of political parties
- Official handles of government departments and agencies like the Ministry of External Affairs, the Reserve Bank of India, etc.

For the US, our definition included:

- President, Vice-President, Governors, Lt-Governors, Mayors and other officials heading the executive branch of government at different levels
- Members of the House of Representatives and Senate of the US Congress, state legislatures and city councils,
- Unelected politicians such as party members (eg. office bearers of the Democratic National Committee (DNC) or the Republican National Committee (RNC)), campaign staff (eg. communications director of presidential/gubernatorial campaigns), state, city and county-level party units
- Official party handles at the national, state and county levels
- SuperPAC handles, college and student wings of political parties
- Official handles of government departments and agencies

We have built NivaDuck - the Marathi word for selector - an ML-based classification pipeline that produces a highly precise set of politicians on Twitter. It consists of three phases (see fig. 1), which included data gathering, two stages of classification and manual verification. As NivaDuck's code is proprietary, we are currently working to secure permissions to release it to the benefit of the wider research community. We used NivaDuck to collect accounts of politicians and pulled their tweets with Twitter's public API. This database of tweets - consisting in excess of 80 million tweets of Indian politicians and more than 40 million tweets of US politicians - is called PoliTwictionary. While this paper focuses only on NivaDuck, we have used our database of tweets for myriad analyses, including the use of hashtags in election campaigns, role of celebrities in shaping political discourse [41, 49], use of extreme speech [66], Twitter trending topics [42], centrality of leaders in online campaigns [43], topical preferences of politicians' tweets [67], and the similarity of their tweets.

The paper is organised as follows. Section 2 refers to prior research that motivated our work. Section 3 covers the building blocks

[1]Lok Sabha is the directly elected lower house of the bicameral Indian parliament; Rajya Sabha is the upper house, with members nominated by state legislatures

of NivaDuck - including the method of data collection, choice of features and classification model, and manual verification. Section 4 details two experiments to estimate the completeness and scope of politicians caught by NivaDuck. In section 5 we conclude with our findings and their relevance to the wider social media research community. Finally, section 6 includes the limitations of our methodology and the resulting dataset, and our plans to expand NivaDuck's capabilities in the future.

## 2 RELATED WORK

Our work is related to research on the aggregate impact of political actors' action on social media discourse. Prior work on social media communication of politicians has used a variety of methods to source and aggregate data from Twitter or Facebook accounts. Shapiro et al. [76] used data from a list of 202 Korean politicians compiled from the official page of the Korean National Assembly [2]. In their study of members of the Swiss Federal Parliament, Rauchfelisch et al. [73] collected 81 Twitter handles manually from parliamentary records of the names of politicians. Hemphill et al. [36] used a data set of Twitter handles of 380 members of the US Congress to compare politicians' online behaviour by gender and party. These methods limit us to only published datasets of key national politicians, such as parliamentarians, and cannot be extended to cover local, unelected leaders or grassroots activists that are not well documented.

Grant et al. [31] used an expansive definition of politicians, including federal and state representatives and declared candidates for those positions in Australia, for which they searched politicians' websites and conducted advanced Google searches for networks of known politicians. They removed obvious fakes and clarified the ones that had low information. However, it too was limited to relatively well-known elected or contesting politicians and required significant human effort to manual collect and verify the list of 152 Australian politicians. We observe that studies which provide deep understanding of the political social media rely upon well documented data sets, often published by prior research.

Building such a database manually poses multiple challenges for large electoral systems. It is an expensive and time consuming process to gather a large corpus of such accounts, and political status and affiliation are highly fluid, especially in parliamentary systems. Researchers have devoted significant effort using myriad methods to classify social media users into pertinent categories. In their study [29] on categorising Facebook users according to the 'Big Five' [16] personality traits, Golbeck et. al utilised a combination of structural, profile based, linguistic and statistical features. Lima et al. [51] used the same Big Five model to classify users based on their tweets of popular Brazilian TV shows. Tang et al. featurized [83] heterogeneous networks of social media users to improve classification performance. Bergsma et al. [7] utilised the names of users and their mutual interactions to predict ethnicity and gender in a scalable framework. Singh et al. [77] presented a method to identify spammers of pornography on Twitter using their tweet as a feature vector.

Rao et al. [72] explored features such as n-gram models of user's tweets, their network statistics, socio-linguistic attributes like emoticons, special characters and symbols, etc., and communication patterns (like rate of retweeting) to infer latent information such as the user's gender, age, regional origins and their political orientation. Pennacchiotti et al. built [70] upon their work and assessed the relative importance of features like profile information, tweeting behaviour, linguistic aspects of tweets and the social network of one's Twitter activity (friend and retweets networks) in predicting the user's political orientation [69] and ethnicity. Based on their findings regarding the usefulness of profile description text, we expanded our feature set to include tweets in a second classifier.

## 3 NIVADUCK'S PIPELINE

Figure 1 details the procedure we used to build our annotated dataset of politicians' Twitter accounts.

The pipeline begins with the `seed` data i.e. a manually curated list of major political figures on Indian and American Twitter. This data is used in two ways: as the positive class of the training set and for 'snow-balling' - using the Twitter networks of known politicians to identify new ones for the `inference` set. We snow-balled on the friend network, follower network and the list network of the seed data to build the inference set.

The `inference` set is passed through the pre-processing stage, removing accounts that cannot be classified by NivaDuck such as those with fewer than 50 tweets and those already present in our database. The remaining accounts are then featurized and fed to the classifier for prediction. In the classification pipeline, the primary classifier uses profile description text to produce a high-recall set to the secondary classifier. In this stage, the classifier produces a high-precision set of politicians. The predicted political handles proceed to the final stage - manual verification. In this stage, we verify each account to weed out false positives i.e. accounts wrongly labeled as politicians by virtue of similarity of their description and tweets with known politicians. We also added useful annotations such as party affiliation, state in which they operate and level in the party's scalar chain. All false positives are fed back to NivaDuck's training set as negative class samples for learning. A random sample of true positives identified in this step are also included in the classifier's `training` set. Finally, all true positives are added to the `database` with suitable metadata.

### 3.1 Classification step

For the United States, we built our labelled sample set using the Twitter handles of the members of the US Congress. To this set, we added politicians and non-politicians from the a random sample of 1000 accounts from the list network [3] of the Democratic party's congressional handle @HouseDemocrats. Our final labelled set consisted of 1500 unique samples, including 648 politicians.

Our sample set for India included 3283 labeled samples that were collected manually. These consisted of 1667 handles of politicians at the national (excluding members of parliament), state and local levels from 42 political parties. The distribution of accounts for the

[2] http://korea.assembly.go.kr/

[3] The list network of a Twitter handle refers to all accounts listed with that account in Twitter `lists` by regular Twitter users: `https://help.twitter.com/en/using-twitter/twitter-lists`

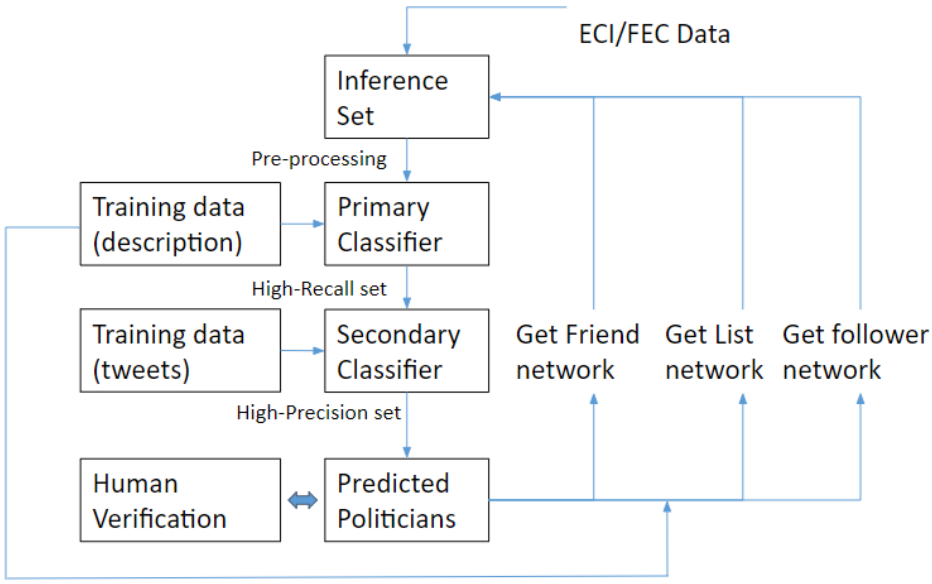


**Figure 1: Classification pipeline to build the datatset of politicians**

top 20 parties was as follows: BJP[4]-413, INC[5]-348, AAP-76, SP-69, DMK[6]-65, AIADMK-61, SS-57, TRS-56, AITC-53, NCP-52, AIMIM-51, CPIM-51, YSRCP-48, BJD-48, TDP-48, JDS-14, RJD-13, GOV-13 [7], SAD-13, JKNC-12. The negative set consisted of non-political handles from among Twitter friends of politicians, of politically active members of the public collected using political hashtags, supporters of major Indian political figures and parties (marked by party names such as 'BJP', 'INC', 'DMK', etc.), and mix of top retweeted handles in India.

We used these labeled samples to train NivaDuck's calssifiers, to be used in the pipeline to precisely identify new politicians at scale.

## 3.2 Choice of features and classifier

For the primary classifier, we chose description text as the feature for classification. Tables 1 and 2 include the Twitter descriptions of a few samples from our labeled dataset for India and the US. To preserve the indic keywords described above, we did not stem the tokens in the featurization stage for the India dataset. Next, we expanded the feature vector of unigrams to include bigrams and trigrams and used NLTK's `collocations` [9, 53] toolkit to extract these features. We ordered them by the `likelihood_ratio` reported by the `BigramCollocationsFinder` (and `TrigramCollocationsFinder`), selected the top 1000 of each type and then calculated the significance $S_{bigram}$ and $S_{trigram}$ as follows:

$$S_{bigram} = \log\left[\frac{¶(bigram|class == 1)}{¶(bigram|class == 0)}\right] \quad (1)$$

$$S_{trigram} = \log\left[\frac{¶(trigram|class == 1)}{¶(trigram|class == 0)}\right] \quad (2)$$

where $¶(B|class == 1)$ refers to the probability of bigram B occurring in the positive class samples of the training set. The value of S is positive for all bigrams that occur much more in the positive class samples and negative for those that are frequent in the negative class samples. We selected the two hundred bigrams with the most positive and most negative significance values, giving us a set of 400 bigram features that differentiated the two classes. On repeating the procedure for trigrams, we found less than five trigram samples that had negative significance values and most had very small positive values. We therefore selected only the top 80 trigrams with positive S scores.

To select the primary classifier's unigram features, we used a TF-IDF [68] vectorizer. We ordered them by their inverse document frequency and omitted the top 50 unigrams. This step produced 1125 and 918 unigram features for the India and USA training data respectively. We combined these unigram features with the bigrams and trigrams and fit a TF-IDF vectorizer to the training data using this vocabulary.

We selected a `LogisticRegression` classifier to preserve the model's simplicity and to have a transparent process that allowed us to measure the relative weights of features. We used `GridSearchCV` to optimize the primary classifier's performance. Moreover, we used the precision recall curve of the cross-validation set to determine the correct threshold for classification to produce a high recall prediction from the primary classifier. The value of $T$ was set to the threshold corresponding to the highest precision value for a

[4]BJP - Bhartiya Janata Party (Indian People's Party)
[5]INC - Indian National Congress
[6]DMK - Dravida Muneitra Kazhagam
[7]Official government handles like @PMOIndia, @CPMumbaiPolice, @SmritiIraniOffc were marked as 'GOV' for party

minimum threshold of 95 percent recall [8]. The value of $T$ were 0.22 and 0.13 respectively for India and USA.

Table 3 details the training, cross-validation and testing performance with these thresholds. Table **??** shows the 20 most positive and negative features for both training sets. We used this configuration with the `inference` set to feed accounts to the secondary classifier.

**Table 1: A sample of ten politicians from the training set and their Twitter descriptions (bold text indicates keywords for classification, underlined words are specific to Indian politics)**

| screen_name | description |
|---|---|
| RahulGandhi | This is the official account of Rahul Gandhi \| **Member of Parliament** \| **President, Indian National Congress** |
| narendramodi | **Prime Minister** of India |
| AjayChoudhariSS | Official Account of Mr. Ajay Chourdhari **MLA**, Shivadi **Vidhansabha**(Lalgbaug, Parel, Sewri) |
| mieknathshinde | **Cabinet Minister** for Public Works (Undertaking) **Government** of Maharashtra \| **MLA** - Kopri-Pachpakhadi **Vidhan Sabha** \| Guardian **Minister** - Thane \| **ShivSena** |
| Pradeep_Behera7 | **Youth Congress Vice President** , Rourkela |
| SureshHalwankar | **Member Of Legislative Assambly**, Maharashtra \| **General Secretary, BJP** Maharashtra \| Member, Public Undertaking Committee \| **Ex President, BJP Kolhapur Dist.** |
| KInaochaDevi | **Zilla Parishad Member**. **President, Mahila Morcha, BJP** Manipur **Pradesh**. |
| dineshoraon6 | **Speaker, Jharkhand Legislative Assembly** |
| DKSureshINC | Indian **Member of Parliament** \| Represents the Bangalore Rural **constituency** of Karnataka \| **Indian National Congress** |
| Ranjeet4India | **Member of Parliament** from Supaul, Secretary - **AICC, National Spokesperson - Congress** |

For the secondary classifier, we featurized the tweets of politicians to produce a highly precise set of potential politicians. We selected tweets from specific periods - Jan-May 2019 for India and April-Aug 2019 for the USA - to account for temporal variations in tweet content and varying frequency of tweeting between politicians. We concatenated all tweets of a given account and used Google's Universal Sentence Encoder [14] to represent them as 512-size word embedding. These feature vectors are fed to a `LogisticRegression` classifier, optimized using `GridSearchCV` with biased weights to prioritize precision over recall. Table 4 shows the performance metrics of the secondary classifier.

[8]For the India dataset, the minimum recall requirement was increased to 98% to account for the diversity of the training data

**Table 2: A sample of ten politicians from the US training set and their Twitter descriptions (bold text indicates keywords for classification, underlined words are specific to US politics)**

| screen_name | description |
|---|---|
| ColoSenGOP | The official Twitter page for the Colorado **Senate Republicans**. |
| RepThompson | Rep. Mike Thompson represents California's 5th **Congressional District** and chairs the **House Gun Violence Prevention Task Force**. |
| RepLBR | Official Twitter page for **U.S. Representative** Lisa Blunt Rochester (D-DE). |
| ohiogop | The official Twitter page of the **Ohio Republican Party**. |
| AGKarlRacine | Official Twitter account of the Office of the **Attorney General** for the **District of Columbia**.\nFacebook: https://t.co/Ge4fiCXEAQ |
| ClarkJolley | Member, **Oklahoma Tax Commission** |
| BankingGOP | **Senate** Banking, Housing & Urban Affairs **Republicans** |
| sethmoulton | Father, husband, Marine, **Congressman**, and **candidate** for **President** of the United States. |
| RepBillJohnson | Proudly representing **#Ohio's 6th Congressional District**. Energy Enthusiast. Veteran. Husband. Father. Grandfather. |
| KansasDems | Powered by, fighting for working **Kansans**. |

**Table 3: Performance metrics of NivaDuck's primary classifier after cross-validation**

India

| | dataset | f1 score | precision | recall | accuracy |
|---|---|---|---|---|---|
| 1 | Training set | 0.76 | 0.61 | 0.99 | 0.68 |
| 2 | Cross-Validation set | 0.77 | 0.63 | 0.97 | 0.69 |
| 3 | Test set | 0.75 | 0.61 | 0.98 | 0.67 |

USA

| | dataset | f1 score | precision | recall | accuracy |
|---|---|---|---|---|---|
| 1 | Training set | 0.90 | 0.82 | 0.99 | 0.87 |
| 2 | Cross-Validation set | 0.87 | 0.80 | 0.97 | 0.84 |
| 3 | Test set | 0.88 | 0.80 | 0.97 | 0.85 |

## 4 COMPLETENESS AND SCOPE OF NIVADUCK

To estimate the completeness and scope of accounts caught by NivaDuck, we ran two experiments. For the first, we manually gathered the Twitter handles of elected members of the Indian parliament. For the USA, we compiled the list of all members of the California State Assembly and Senate as well as candidates that ran for those positions during the 2018 elections. We repeated the procedure for the San Diego city council. Our goal was to estimate

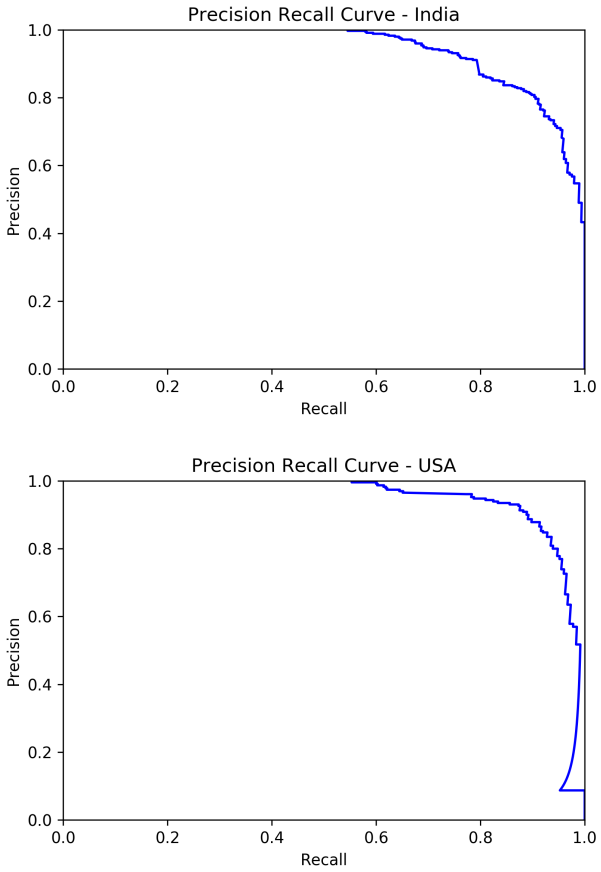


**Figure 2: Preicision Recall Curves of NivaDuck's primary classifier**

**Table 4: Performance metrics of NivaDuck's secondary classifier after cross-validation**

| | India | | | | |
|---|---|---|---|---|---|
| | dataset | f1 score | precision | recall | accuracy |
| 1 | Training+Cross Validation | 0.80 | 0.88 | 0.74 | 0.82 |
| 2 | Test set | 0.76 | 0.87 | 0.68 | 0.74 |
| | USA | | | | |
| | dataset | f1 score | precision | recall | accuracy |
| 1 | Training+Cross Validation | 0.96 | 0.98 | 0.94 | 0.95 |
| 2 | Test set | 0.86 | 0.92 | 0.81 | 0.86 |

what percentage of these handles could be caught by NivaDuck using the seed data we had provided.

To build the ground truth for this exercise, we used the list of elected representatives published by the Election Commission of India (ECI) and manually looked up the name of each MP on Twitter to find their Twitter handle. We found 424 Twitter handles for 545 MPs of the Lok Sabha. For the California set, we used data published by the Federal Election Commission (FEC) and state level official sources [9] and candidate lists as per Ballot-o-pedia [10]. We repeated the procedure for San Diego county. For each politician, we recorded their personal, official and campaign accounts, if any. As an example, Democratic Assemblywoman Cecilia Aguiar-Curry has two accounts - official handle @AsmAguiarCurry and personal handle @CeciliaAD4. Similarly, Republican State Senate candidate Rex Hime had one official handle @RexHime and one campaign handle @HimeForSenate during the 2018 elections. In all, we found 153 active handles for 171 accounts we looked up. Most CA politicians had only one Twitter handle. We repeated the procedure for San Diego County representatives and compiled 61 accounts for 56 politicians.

The rest of the experiment was set up as follows. To find the aforementioned handles automatically, NivaDuck would snowball on the seed data through Twitter friend and follower networks. It would then feed all accounts so found to its two-stage classification pipeline. The accounts found to be politicians would then indicate NivaDuck's ability to find elected representatives in India and the US.

For the India set, NivaDuck found 421 MPs using the friend and follower network, caught 401 of these handles in its high-recall primary classifier and predicted 284 MPs as potential politicians in its high-precision secondary classifier. This yields a overall recall score of 65 percent for the Indian MPs dataset. Using the same procedure for the US, we found that NivaDuck caught 68 percent of California State handles and 53 percent of San Diego County handles. The low recall scores were expected as NivaDuck's secondary classifier is trained to produce a highly precise set of politicians and filters out any accounts whose tweets are dissimilar from those of politicians in the training data. This is especially punitive towards local, county-level politicians, as their tweets are expected to focus on substantially different issues than national politicians in the training data. That NivaDuck yields the lowest recall on the San Diego county set corroborates this hypothesis.

The goal of the second experiment was to gauge the scope and diversity of accounts that NivaDuck can find. We ran the pipeline for the entire set of friend, follower and list network of our seed sets. We then sampled one thousand handles of predicted politicians. We verified that these accounts were politicians and annotated them by party (for India), level (for the USA) and state (both). Being multiparty parliamentary democracy, we prioritized party annotations for India. Of these samples, 913 accounts were verified to be politicians for the US set while 867 were true positives for the India set.

Notably, NivaDuck successfully found a large and diverse set of handles in the USA, with all 50 states, the District of Columbia and Puerto Rico being represented. While the distribution of handles across states is not directly proportional to the number of elected federal representatives or population, we surmise that our oversampling of accounts from states like Ohio, North Carolina and

[9] State Senators: https://www.senate.ca.gov/senators, State Assembly: https://www.assembly.ca.gov/assemblymembers
[10] https://ballotpedia.org/Main_Page

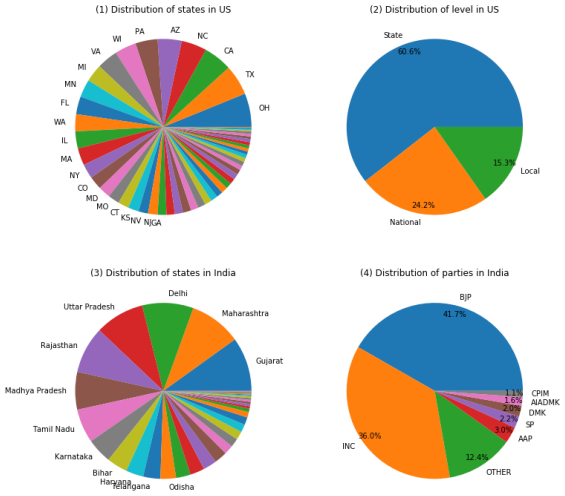


**Figure 3: Scope of political handles found by NivaDuck**

Pennsylvania may be attributed to these being swing states. In India, NivaDuck found politicians across all 28 states and 9 union territories as well as 'Overseas' accounts such as @AAPUSAOfficial, @INCOverseas and @ysrcp_australia, operated by Indian diasporas worldwide. In India, PM Modi's home state of Gujarat leads in number of politicians, majorly due to a disproportionate number of BJP handles. States like Maharashtra (with big cities like Mumbai and Pune) and Delhi (the national capital) follow. We find a diverse set of northern (Rajasthan, Uttar Pradesh, etc.) and southern (Tamil Nadu, Karnataka, etc.) states among the top 12 shown in figure 3. This wide scope affords us the ability to build a comprehensive, if not complete, archive of political Twitter accounts in a given country. The figure excludes the false positives found in the verification stage - 87 (8.7%) for the USA set and 133 (13.3%) for the India set.

Appendix A includes links to our public GitHub repository that provides the list of all politicians by country, with metadata such as state and party affiliation. We also plan to differentiate elected and unelected leaders in our dataset and will continue to update this information while adding new politicians in the future.

## 5 DISCUSSION

We have presented a precise and scalable framework to identify large datasets of political figures, with minimal time-space constraints. NivaDuck has so far identified 18500+ Indian politicians and 8000+ US politicians at the national, state and district levels. We reported on the completeness and scope of our methodology.

We have used this dataset for myriad studies. We accomplished first large scale study on topical partisan preferences of over 7400 Indian politicians during the 2019 general elections. We have also analysed the effect of using extreme speech - the overlapping use of humor, satire, insults, hate speech - on retweets earned by politicians in India. Moreover, we found that politicians from the Hindi-speaking states and the southern state of Tamil Nadu are highly unlikely to use English whereas those from the North Eastern states preferred it over their local languages. Regional parties like the DMK, AIADMK, SP, BSP, JDU, Shiv Sena, NCP, YSRCP and TDP are unlikely to use English whereas parties in Jammu and Kashmir and Punjab, and national government accounts had strong inclination to use the language. Finally, we have used PoliTwictionary to study the centrality of mentions of leaders in party's election campaigns, the dominance of parties in Twitter trending topics and interaction between politicians and celebrities on Twitter[11]

These analyses were made possible by NivaDuck's ability to build a diverse corpus of politicians to populate PoliTwictionary.

Our work carries significant value to the research community. First, it can provide well annotated data on a large scale to social media archives like SOMAR, built by Hemphill et al. [35]. This database would also allow extensions of work on politicians Twitter content, such as their interactivity with constituents (or lack thereof) [61] [60] [38], topic modelling of their tweets [34, 36] and their tweets' correlation with mass media reporting [75]. Second, using categories of party, state and level, one can study homophily [32] among politicians of a particular group, degree of polarization [10, 74] between groups and their activity during major political events such as political debates [52, 54]. Third, politicians' retweet network and mention frequency can be studied to assess patronage of specific leaders, degree of centrality in parties' internal structure and favoritism by senior leaders towards their peers and supporters. Fourth, there are also content-specific lessons we can draw by creating such lists. Large lists can be used to study trolling behaviour,

[11]These works are in submission or have not been cited to preserve anonymity

misinformation provenance and collusion among party functionaries, in line with prior studies on generic Twitter users [26]. This system can also be useful in temporal analysis of politicians' roles and positions, helping us track the pathways politicians followed to reach higher office.

## 6 LIMITATIONS AND FUTURE WORK

In this paper, we have documented our efforts towards building a machine learning-based methodology to compile a large dataset of politicians for the two largest democracies of the world - India and the USA. While we tried to achieve the task purely through automated scraping of accounts using the Twitter API and ML-based classification using NivaDuck's classifiers, the results were underwhelming. Notably, as NivaDuck looks for politicians among the Twitter networks of politicians in the seed set, several key politicians from smaller parties that are not linked to this set may escape its reach. Moreover, given that its secondary classifier is trained for high-precision, NivaDuck yields a high false negative rate for elected state and local politicians. In addition, while we did not find major regional imbalances for the USA, the multilingual polity in India can exacerbate bias against smaller regional parties if the seed set is not sufficiently large and diverse. Also, due to the inclusion of tweets as a feature in the classification pipeline, politicians who are inactive on Twitter (i.e. tweet infrequently or haven't tweeted in a long time) may drop through the cracks regardless of their relevance in the real world. This may lead to the dataset missing important politicians. Finally, the pipeline needs to be run repeatedly to account for changes in status of politicians, such as elections or politicians changing their party - a common phenomenon in India.

To overcome these problems, we manually added members of parliament who were absent from the list. For the India set, we also manually verified all friends of politicians that had verified status or had at least 150 politicians from our dataset following them, and complemented the dataset NivaDuck had built. We found 1002 handles from amongst friends of over 17500+ politicians. Each of these accounts was then verified manually and annotated with labels for party, state and level. We plan to repeat the same process for the USA dataset.

Moving ahead, we intend to complement NivaDuck's dataset with attributes like gender, ethnicity, party affiliation and geographical location, obtained by cross-referencing our data with reliable public sources of such information. These would be verified manually for accuracy. We also plan to deploy NivaDuck in other countries, especially in the Global South. Our eventual goal is to build and maintain an archive of politicians tweets, their profile characteristics and relevant demographic and political metadata.

## ACKNOWLEDGMENTS

We would like to thank Lia Bozarth for providing the initial seed set of Indian politicians. We appreciate the efforts of Faisal Lalani, Azhagu Meena, Zainab Akbar, Ramaravind Kommiya Mothilal, and Gopal Srinivasa for their contributions to verifying and annotating the dataset. We are also grateful to Ashwin Rajadesingan for his insightful review of the work.

## REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.
[2] Nezar AlSayyad and Muna Guvenc. 2015. Virtual uprisings: On the interaction of new social media, traditional media coverage and urban space during the 'Arab Spring'. *Urban Studies* 52, 11 (2015), 2018–2034.
[3] Chinmayi Arun. 2019. On WhatsApp, Rumours, and Lynchings. *Economic & Political Weekly* 54, 6 (2019), 31.
[4] Kai Arzheimer. 2015. The afd: Finally a successful right-wing populist eurosceptic party for germany? *West European Politics* 38, 3 (2015), 535–556.
[5] W Lance Bennett. 2012. The personalization of politics: Political identity, social media, and changing patterns of participation. *The ANNALS of the American Academy of Political and Social Science* 644, 1 (2012), 20–39.
[6] Annie Benveniste and Etienne Pingaud. 2016. Far-right movements in France: The principal role of Front National and the Rise of Islamophobia. In *The Rise of the Far Right in Europe*. Springer, 55–79.
[7] Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. 2013. Broadly improving user classification via communication-based name and location clustering on twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1010–1019.
[8] Mohsin Alam Bhat. [n.d.]. How Mob Violence and Hate Crimes Are Linked to Social Vulnerability. http://www.jgls.edu.in/article/how-mob-violence-and-hate-crimes-are-linked-social-vulnerability
[9] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.
[10] Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and Network Dynamics Behind Egyptian Political Polarization on Twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. ACM, New York, NY, USA, 700–711. https://doi.org/10.1145/2675133.2675163
[11] Shelley Boulianne. 2015. Social media use and participation: A meta-analysis of current research. *Information, communication & society* 18, 5 (2015), 524–538.
[12] Manuela Caiani and Linda Parenti. 2009. The dark side of the web: Italian right-wing extremist groups and the Internet. *South European Society and Politics* 14, 3 (2009), 273–294.
[13] Guido Caldarelli, Alessandro Chessa, Fabio Pammolli, Gabriele Pompa, Michelangelo Puliga, Massimo Riccaboni, and Gianni Riotta. 2014. A multi-level geographical study of Italian political elections from Twitter data. *PloS one* 9, 5 (2014), e95809.
[14] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).
[15] Andrea Ceron, Luigi Curini, Stefano M Iacus, and Giuseppe Porro. 2014. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society* 16, 2 (2014), 340–358.
[16] Deborah A Cobb-Clark and Stefanie Schurer. 2012. The stability of big-five personality traits. *Economics Letters* 115, 1 (2012), 11–15.
[17] Derrick L Cogburn and Fatima K Espinoza-Vasquez. 2011. From networked nominee to networked nation: Examining the impact of Web 2.0 and social media on political participation and civic engagement in the 2008 Obama campaign. *Journal of Political Marketing* 10, 1-2 (2011), 189–213.
[18] Shoaib Daniyal. [n.d.]. The Modi Years: What has fuelled rising mob violence in India? https://scroll.in/article/912533/the-modi-years-what-has-fuelled-rising-mob-violence-in-india
[19] Jenny L Davis, Tony P Love, and Gemma Killen. 2018. Seriously funny: The political work of humor on social media. *new media & society* 20, 10 (2018), 3898–3916.
[20] Joseph DiGrazia, Karissa McKelvey, Johan Bollen, and Fabio Rojas. 2013. More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one* 8, 11 (2013), e79449.
[21] Martin Dolezal. 2015. Online campaigning by Austrian political candidates: Determinants of using personal websites, Facebook, and Twitter. *Policy & Internet* 7, 1 (2015), 103–119.
[22] Mattias Ekman. 2015. Online Islamophobia and the politics of fear: manufacturing the green scare. *Ethnic and Racial Studies* 38, 11 (2015), 1986–2002. https://doi.org/10.1080/01419870.2015.1021264 arXiv:https://doi.org/10.1080/01419870.2015.1021264
[23] Sven Engesser, Nicole Ernst, Frank Esser, and Florin Büchel. 2017. Populism and social media: How politicians spread a fragmented ideology. *Information, communication & society* 20, 8 (2017), 1109–1126.
[24] Gunn Enli. 2017. Twitter as arena for the authentic outsider: exploring the social media campaigns of Trump and Clinton in the 2016 US presidential election. *European journal of communication* 32, 1 (2017), 50–61.
[25] Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. (2017).

[26] Claudia Flores-Saviaga, Brian C. Keegan, and Saiph Savage. 2018. Mobilizing the Trump Train: Understanding Collective Action in a Political Trolling Community. *CoRR* abs/1806.00429 (2018). arXiv:1806.00429 http://arxiv.org/abs/1806.00429

[27] Peter L Francia. 2018. Free media and Twitter in the 2016 presidential election: The unconventional campaign of Donald Trump. *Social Science Computer Review* 36, 4 (2018), 440–455.

[28] Paolo Gerbaudo. 2018. Social media and populism: an elective affinity? *Media, Culture & Society* 40, 5 (2018), 745–753.

[29] Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting Personality with Social Media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI EA '11)*. ACM, New York, NY, USA, 253–262. https://doi.org/10.1145/1979742.1979614

[30] Annie Gowen. [n.d.]. As mob lynchings fueled by WhatsApp messages sweep India, authorities struggle to combat fake news. https://www.washingtonpost.com/world/asia_pacific/as-mob-lynchings-fueled-by-whatsapp-sweep-india-authorities-\struggle-to-combat-fake-news/2018/07/02/683a1578-7bba-11e8-ac4e-421ef7165923_story.html?noredirect=on&utm_term=.8a0a0f542b7c

[31] Will J Grant, Brenda Moon, and Janie Busby Grant. 2010. Digital dialogue? Australian politicians' use of the social network tool Twitter. *Australian Journal of Political Science* 45, 4 (2010), 579–604.

[32] Catherine Grevet, Loren G. Terveen, and Eric Gilbert. 2014. Managing Political Differences in Social Media. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW '14)*. ACM, New York, NY, USA, 1400–1408. https://doi.org/10.1145/2531602.2531676

[33] Aniko Hannak, Drew Margolin, B Keegan, and I Weber. 2014. Get Back! You don't know me like that: The social mediation of fact checking interventions in twitter conversations. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* (01 2014), 187–196.

[34] Libby Hemphill, Aron Culotta, and Matthew Heston. 2013. Framing in Social Media: How the US Congress uses Twitter hashtags to frame political issues. *Available at SSRN 2317335* (2013).

[35] Libby Hemphill, Susan H Leonard, and Margaret Hedstrom. 2018. Developing a Social Media Archive at ICPSR. (2018).

[36] Libby Hemphill, Jahna Otterbacher, and Matthew Shapiro. 2013. What's congress doing on twitter?. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 877–886.

[37] Libby Hemphill and Andrew J Roback. 2014. Tweet acts: how constituents lobby congress via Twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 1200–1210.

[38] Sounman Hong and Daniel Nadler. 2012. Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on candidate salience. *Government Information Quarterly* 29, 4 (2012), 455 – 461. https://doi.org/10.1016/j.giq.2012.06.004 Social Media in Government - Selections from the 12th Annual International Conference on Digital Government Research (dg.o2011).

[39] Sam Graham-Felsen Hughes, Kate Allbright-Hannah, Scott Goodstein, Steve Grove, Randi Zuckerberg, Chloe Sladden, and Brittany Bohnet. 2010. Obama and the power of social media and technology. *The European Business Review (May-June 2010)* (2010), 16–21.

[40] Maurice Jakesch and Koren. 2018. Moran and Evtushenko. (2018). http://dx.doi.org/10.2139/ssrn.3306403

[41] Anmol Panda Joyojeet Pal. 2019. Narendra Modi matinee show: Inside India's celeb Twitter. *Livemint* (2019). https://www.livemint.com/politics/news/narendra-modi-matinee-show-inside-india-s-celeb-twitter-1548616373119.html

[42] Faisal Lalani Joyojeet Pal, Anmol Panda. 2019. How #BJP fused with #StrongIndia in 2019. *Livemint* (2019). https://www.livemint.com/elections/lok-sabha-elections/how-bjp-fused-with-strongindia-in-2019-1557414405626.html

[43] Ramgopal Chandrasekaran Joyojeet Pal, Anmol Panda. 2019. On Twitter, Sachin Pilot is CM and Shivraj trumps Modi in MP. *Livemint* (2019). https://www.livemint.com/Politics/OJIYbL2mgZzsMXLyB0GylL/twitter-rajasthan-mp-elections-sachin-pilot-shivraj-singh-ch.html

[44] Andreas Jungherr, Harald Schoen, and Pascal Jürgens. 2015. The mediation of politics through Twitter: An analysis of messages posted during the campaign for the German federal election 2013. *Journal of Computer-Mediated Communication* 21, 1 (2015), 50–68.

[45] James Katz, Michael Barris, and Anshul Jain. 2013. *The social media president: Barack Obama and the politics of digital engagement*. Springer.

[46] Douglas Kellner. 2019. Trump's War Against the Media, Fake News, and (A) Social Media. In *Trump's Media War*. Springer, 47–67.

[47] Habibul Haque Khondker. 2011. Role of the new media in the Arab Spring. *Globalizations* 8, 5 (2011), 675–679.

[48] Daniel Kreiss. 2012. Acting in the public sphere: The 2008 Obama campaign's strategic use of new media to shape narratives of the presidential race. In *Media, movements, and political change*. Emerald Group Publishing Limited, 195–223.

[49] Faisal M Lalani, Ramaravind Kommiya Mothilal, and Joyojeet Pal. 2019. The Appeal of Influencers to the Social Media Outreach of Indian Politicians. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 267–271.

[50] Abbey Levenshus. 2010. Online relationship management in a presidential campaign: A case study of the Obama campaign's management of its internet-integrated grassroots effort. *Journal of Public Relations Research* 22, 3 (2010), 313–335.

[51] Ana Carolina ES Lima and Leandro Nunes De Castro. 2014. A multi-label, semi-supervised classification approach applied to personality prediction in social media. *Neural Networks* 58 (2014), 122–130.

[52] Yu-Ru Lin, Brian Keegan, Drew Margolin, and David Lazer. 2014. Rising Tides or Rising Stars?: Dynamics of Shared Attention on Twitter during Media Events. *PLOS ONE* 9, 5 (05 2014), 1–12. https://doi.org/10.1371/journal.pone.0094093

[53] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1* (Philadelphia, Pennsylvania) *(ETMTNLP '02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 63–70. https://doi.org/10.3115/1118108.1118117

[54] Misa T. Maruyama, Scott P. Robertson, Sara K. Douglas, Bryan C. Semaan, and Heather A. Faucett. 2014. Hybrid Media Consumption: How Tweeting During a Televised Political Debate Influences the Vote Decision. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW '14)*. ACM, New York, NY, USA, 1422–1432. https://doi.org/10.1145/2531602.2531719

[55] Emily Metzgar and Albert Maruggi. 2009. Social media and the 2008 US presidential election. *Journal of New Communications Research* 4, 1 (2009).

[56] Paul Mihailidis and Samantha Viotty. 2017. Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society. *American Behavioral Scientist* 61, 4 (2017), 441–454.

[57] Sriram Mohan. 2015. Locating the "Internet Hindu" Political Speech and Performance in Indian Cyberspace. *Television & New Media* 16, 4 (2015), 339–345.

[58] Chandan Nandy. [n.d.]. Political Silence Over Lynchings is Sanction by Another Means. https://www.newsclick.in/political-silence-over-lynchings-sanction-another-means

[59] Brian L Ott. 2017. The age of Twitter: Donald J. Trump and the politics of debasement. *Critical Studies in Media Communication* 34, 1 (2017), 59–68.

[60] Jahna Otterbacher, Libby Hemphill, and Matthew A Shapiro. 2012. Tweeting vertically? Elected officials' interactions with citizens on Twitter. In *CeDEM (Conference for E-Democracy and Open Government) Asia 2012*.

[61] Jahna Otterbacher, M Shapiro, and Libby Hemphill. 2013. Interacting or just acting. *Journal of Contemporary Eastern Asia* 12, 1 (2013), 5–20.

[62] Mustafa Oz, Pei Zheng, and Gina Masullo Chen. 2018. Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *new media & society* 20, 9 (2018), 3400–3419.

[63] Joyojeet Pal. 2015. Banalities turned viral: Narendra Modi and the political tweet. *Television & New Media* 16, 4 (2015), 378–387.

[64] Joyojeet Pal, Priyank Chandra, and VG Vinod Vydiswaran. 2016. Twitter and the rebranding of Narendra Modi. *Economic & Political Weekly* 51, 8 (2016), 52–60.

[65] Joyojeet Pal and Andre Gonawela. 2016. Political social media in the global South. In *Conference on e-Business, e-Services and e-Society*. Springer, 587–593.

[66] Anmol Panda, Sunandan Chakraborty, Noopur Raval, Han Zhang, Mugdha Mohapatra, Syeda Zainab Akbar, and Joyojeet Pal. 2020. Affording Extremes: Incivility, Social Media and Democracy in the Indian Context. In *Proceedings of the Eleventh International Conference on Information and Communication Technologies and Development* (Guayaquil, Ecuador) *(ICTD '20)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3392561.3394637

[67] Anmol Panda, Ramaravind Kommiya Mothilal, Monojit Choudhury, Kalika Bali, and Joyojeet Pal. 2020. Topical Focus of Political Campaigns and its Impact: Findings from Politicians' Hashtag Use during the 2019 Indian Elections. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '20)*. Association for Computing Machinery, New York, NY, USA, 14. https://doi.org/10.1145/3392860

[68] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[69] Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 430–438.

[70] Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to twitter user classification. In *Fifth International AAAI Conference on Weblogs and Social Media*.

[71] Nathaniel Persily. 2017. The 2016 US Election: Can democracy survive the internet? *Journal of democracy* 28, 2 (2017), 63–76.

[72] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. ACM, 37–44.

[73] Adrian Rauchfleisch and Julia Metag. 2016. The special case of Switzerland: Swiss politicians on Twitter. *New Media & Society* 18, 10 (2016), 2413–2431.
[74] Bryan C. Semaan, Scott P. Robertson, Sara Douglas, and Misa Maruyama. 2014. Social Media Supporting Political Deliberation Across Multiple Public Spheres: Towards Depolarization. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW '14)*. ACM, New York, NY, USA, 1409–1421. https://doi.org/10.1145/2531602.2531605
[75] Matthew A Shapiro and Libby Hemphill. 2017. Politicians and the policy agenda: Does use of Twitter by the US Congress direct New York Times content? *Policy & internet* 9, 1 (2017), 109–132.
[76] Matthew A Shapiro, Libby Hemphill, Jahna Otterbacher, and Han Woo Park. 2014. Twitter and Political Communication in Korea: Are Members of the Assembly Doing What They Say? *Journal of Asia Pacific Studies* 3, 3 (2014).
[77] Monika Singh, Divya Bansal, and Sanjeev Sofat. 2016. Behavioral analysis and classification of spammers distributing pornographic content in social media. *Social Network Analysis and Mining* 6, 1 (2016), 41.
[78] Dominic Spohr. 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review* 34, 3 (2017), 150–160.
[79] Eswaran Sridharan. 2014. Behind Modi's victory. *Journal of Democracy* 25, 4 (2014), 20–33.
[80] Leo G Stewart, Ahmer Arif, and Kate Starbird. 2018. Examining trolls and polarization with a retweet network. In *Proc. ACM WSDM, Workshop on Misinformation and Misbehavior Mining on the Web*.
[81] Kim Strandberg. 2013. A social media revolution or just a case of history repeating itself? The use of social media in the 2011 Finnish parliamentary elections. *New Media & Society* 15, 8 (2013), 1329–1347.
[82] David Talbot. 2008. How Obama really did it. *Technology Review* 111, 5 (2008), 78–83.
[83] Lei Tang and Huan Liu. 2011. Leveraging social media networks for classification. *Data Mining and Knowledge Discovery* 23, 3 (2011), 447–478.
[84] Andy Tattersall. 2018. In the era of Brexit and fake news, scientists need to embrace social media. *LSE Brexit* (2018).
[85] Sahana Udupa. 2018. Gaali cultures: The politics of abusive exchange on social media. *new media & society* 20, 4 (2018), 1506–1522.
[86] Augusto Valeriani and Cristian Vaccari. 2016. Accidental exposure to politics on social media as online participation equalizer in Germany, Italy, and the United Kingdom. *New Media & Society* 18, 9 (2016), 1857–1874.
[87] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
[88] Jerry Watkins, Larissa Hjorth, and Ilpo Koskinen. 2012. Wising up: Revising mobile media in an age of smartphones. *Continuum* 26, 5 (2012), 665–668.
[89] Barry Wellman, Anabel Quan Haase, James Witte, and Keith Hampton. 2001. Does the Internet increase, decrease, or supplement social capital? Social networks, participation, and community commitment. *American behavioral scientist* 45, 3 (2001), 436–455.
[90] Ruth Wodak and Michał Krzyżanowski. 2017. Right-wing populism in Europe & USA. *Journal of Language and Politics* 16, 4 (2017), 471–484.
[91] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2018. Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the Web. *arXiv preprint arXiv:1801.09288* (2018).

## A ONLINE RESOURCES

We have included the complete dataset of Indian and US politicians compiled using NivaDuck in the GitHub repository at: https://github.com/anmolpanda/NivaDuck_public. The same includes meta data such as state, party, level. We will continue to add more politicians, update the metadata and add new details, such as the whether a politician holds elected office, as it changes with future elections.

## B FEATURES OF NIVADUCK'S PRIMARY CLASSIFIER

| | USA dataset | |
|---|---|---|
| | Positive features | Negative features |
| 1 | representing | trade |
| 2 | district | world |
| 3 | senator | human |
| 4 | governor | de |
| 5 | democratic | conservative |
| 6 | representative | communities |
| 7 | rep | politics |
| 8 | congressman | free |
| 9 | serve | teacher |
| 10 | republican | named |
| 11 | real | lawyer |
| 12 | general | theresistance |
| 13 | party | political |
| 14 | congressional | liberal |
| 15 | chairman | retired |
| 16 | senate | mississippi |
| 17 | serving | louisianas |
| 18 | house | advocacy |
| 19 | congresswoman | student |
| 20 | democrats | writer |

**Table 5: Top 20 positive and negative features of the primary classifier trained on the USA dataset**

| | India dataset | |
|---|---|---|
| | Positive Features | Negative Features |
| 1 | mla | government |
| 2 | bjym | news |
| 3 | president | politics |
| 4 | nsui | india |
| 5 | politician | bharat |
| 6 | congress | public |
| 7 | trs | balochistan |
| 8 | secretary | love |
| 9 | aiadmk | pakistan |
| 10 | handle | editor |
| 11 | party | tweets |
| 12 | state | officer |
| 13 | wing | journalist |
| 14 | tdp | fan |
| 15 | youth | supporter |
| 16 | dmk | hard |
| 17 | committee | parody |
| 18 | samajwadi | new |
| 19 | maharashtra | democratic |
| 20 | sabha | ceo |

**Table 6: Top 20 positive and negative features of the primary classifier trained on the India dataset**