Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

DISMISS: Database of Indian Social Media Influencers on Twitter

Arshia Arya[1*], Soham De[2*], Dibyendu Mishra[1*], Gazal Shekhawat[1*], Ankur Sharma[1], Anmol Panda[1], Faisal Lalani[1], Parantak Singh[1], Ramaravind Kommiya Mothilal[1], Rynaa Grover[1], Sachita Nishal[1], Saloni Dash[1], Shehla Shora[1], Syeda Zainab Akbar[1], Joyojeet Pal[3,1]

[1]Microsoft Research, Bangalore, India
[2]Ashoka University, Sonepat, India
[3]University of Michigan, Ann Arbor, Michigan

Abstract

Databases of highly networked individuals have been indispensable in studying narratives and influence on social media. To support studies on Twitter in India, we present a systematically categorised database of accounts of influence on Twitter in India, identified and annotated through an iterative process of friends, networks, and self-described profile information, verified manually. We built an initial set of accounts based on the friend network of a seed set of accounts based on real-world renown in various fields, and then snowballed "friends of friends" multiple times, and rank ordered individuals based on the number of in-group connections, and overall followers. We then manually classified identified accounts under the categories of entertainment, sports, business, government, institutions, journalism, civil society accounts that have independent standing outside of social media, as well as a category of "digital first" referring to accounts that derive their primary influence from online activity. Overall, we annotated 11580 unique accounts across all categories. The database is useful studying various questions related to the role of influencers in polarisation, misinformation, extreme speech, political discourse etc.

Introduction

In the last two decades of growing social media use, a large number of functions in the public sphere are either driven by, or entirely conducted through online communication. Politics, journalism, brand outreach are among a small number of domain spaces of communications that now have a large online component. Influencers, often individuals or accounts who command a large following online and wield influence either directly or through their ability to get second-order engagement in their extended networks. These influencers play a key role in building or propagating momentum around ideas or products – ranging from kickstarting political campaigns to promoting brands and lifestyle products. A number of occupations, especially in media and politics, increasingly rely on practitioners being successful on social media – journalists and electoral candidates, for instance, can expect their success at work to be either bootstrapped or bolstered by their Twitter, Instagram, Facebook, YouTube, or even TikTok presences. Social media influencers have also contributed to the overall changes in the contemporary information environment, including the growth of misinformation and polarizing bias. Consequently, the role of influencers is critical in analysing questions about the recent and ongoing changes in media ecologies.

By building a database of Indian influencers, We make the following main contributions:

- An annotated set of 11580 unique accounts, manually verified as influencers.
- Annotations with 7 broad categories and 24 subcategories encompassing a variety of influencers, including an annotation of "individual" and "entity" type for each subcategory, usable for various research purposes.
- The methods used in this dataset can be used to build similar datasets for other groups of accounts – arranged by nation, state or domain.
- The first major attempt at making an influencer dataset set in the Global South made available for public use.
- This is the first large dataset, to our knowledge, that allows aggregated analysis on accounts categorized by industry and occupations, covering a vast majority of influencers in public life within an ecosystem.

Our dataset is available on Dataverse (DOI: https://doi.org/10.7910/DVN/BPY2JY)

Related Work

Twitter has been of great interest as a data-source to researchers. The Twitter API provides an easy interface to access user and tweet information. However, critical user attributes, such as geo-location, gender and political leanings are not accurately or reliably represented in the Twitter API. To this end, there have been many attempts at annotating these attributes for users. Cheng, Caverlee, and Lee (2010) proposes a Machine Learning approach based exclusively on the tweets made by a user, in the absence of all geospatial clues. Mahmud, Nichols, and Drews (2021) builds on the prior literature by introducing a hierarchical classifier to improve prediction accuracy.

Other attributes, such as gender, age, ethnicity and political affiliation have also been of recurring interest in the research community. Conover et al. (2011) uses an SVM on a manually annotated set of training labels to predict political

*Indicates equal contributions

affiliation of users based on their tweets and hashtag usage. Pennacchiotti and Popescu (2021) predicts political orientation and ethnicity by leveraging other observable information such as Twitter networks and user behaviour. Al Zamal, Liu, and Ruths (2021) attempts to infer these attributes from the neighbours (friends) of a user - an approach based on similar underlying assumptions as ours.Unfortunately, a recent study by Cohen and Ruths (2021) claims that most attempts at annotating political affiliations using Machine Learning systematically report overoptimistic accuracies (nearly 30% higher than expected) due to the way validation datasets are built.

Data Collection

We started the data collection process with the assumption that a seed set can be built using twitter accounts that major politicians choose to follow, using the logic that politicians will follow other politicians or accounts that have some importance in the public sphere - such as media, journalists, key influencers etc. To do this, we used an existing list of known political accounts in India Panda et al. (2020), culled the list of their friends using the Twitter API, and removed all known politicians from the thus exapnded set. From the remainder, we manually removed non-Indian accounts (i.e. accounts that originated from, or relating to an non-Indian person, defined as someone primarily known for their activity within India).

This process resulted in 100k+ accounts from which we removed accounts that are not followed by at least three users from the initial seed set, in order to eliminate accounts that were highly likely to be one-off friends of individuals on the seed set. This was verified manually. At the end of this process, we were left with approximately 10k Twitter accounts of potential influencers that are highly followed by Indian politicians.

This initial set was coded manually, thereupon we did repeat iterations to find accounts followed by journalists, using the assumption that journalists follow newsmakers. We found after two iterations that while we were able to cover a majority of public figures with active twitter accounts. We did this through an exercise of various team members seeking out entertainers, sportspersons, journalists to check if they made the list. We found that the process biased our sample towards more Hindi-speaking states, since they dominate the national narrative.

To mitigate this and ensure more representation among regional states, we ran the same process with journalists and politicians in regional states to ensure more equitable coverage. This process nonetheless has certain disadvantages – a sampling process starting with a seed set of sportspersons or businesspersons would for instance have a relatively larger set of people in sports or business and so on. We used politicians and journalists since they are generally interested in influential individuals across domains.

The team's contextual knowledge of India is used to ensure that known public figures are included in the sample as far as possible. Our process therefore leans more heavily towards influencers with a general appeal than those who are very specific to a field. For instance, a journalist with as few as 3000 Twitter followers may be included in our list, while a film PR agent with 100k followers may end up excluded if they did not have followers from across domains. We also apply the primary domain of engagement by an individual at any given point. Thus, if a person is primarily known for their sporting activities, but also has business interests (such as cricket players Virat Kohli or MS Dhoni) they are nonetheless considered sportspersons in our sample. Likewise a journalist of repute who has written a book or has a leadership role in a media entity, such as Rajdeep Sardesai or Paranjoy Guha Thakurta, would nonetheless be classified as a journalist rather than as an author or businessperson.

Applications of the Data

Our goal of making this data public is to allow researchers to use a comprehensive list of influencers to study various forms engagement and influence on social media on topics ranging from brand management and political outreach to dangerous speech and disinformation in the Indian context. This database is uniquely flexible as the the categories and subcategories can be expanded/collapsed based on the use case. The methodology can be iteratively repeated to get a highly curated list of required categories. Some applications of the database in past work have been described below:

- Dash et al. (2021b) explore influencer polarisation during political crises in India and find that influencers engage with the controversial topics in a partisan manner, with polarized influencers being rewarded with increased retweeting and following. They also observe that specific groups of influencers, particularly fan accounts and platform celebrities consistently engage in polarizing behavior online, thereby underscoring the importance of influencers in political discourse on social media platforms.
- In another study that explored the manifestations of extreme speech through a case study of violent protests and policing in the city of Bangalore, provoked by a derogatory Facebook post, Dash et al. (2021c) found that influential accounts were central in manipulating the discourse surrounding the incident. The dominant narratives that were propagated, employed whataboutism to deflect attention from the triggering post and serve as breeding grounds for religion-based extreme speech.
- The role of influential accounts in disseminating dangerous speech on social media was studied in (Dash et al. 2021a), where they identified dangerous speech by influential accounts on Twitter in India around three key events, examining both the language and networks of messaging that condoned or actively promoted violence against vulnerable groups. They found that dangerous users are more active on Twitter as compared to other users as well as most influential in the network, and act as "broadcasters" in the network, where they are best positioned to spearhead the rapid dissemination of dangerous speech across the platform.
- Mishra, Sen, and Pal (2021) uses our dataset as a seed set for sportspersons and goes over multiple iterations of the workflow to obtain a list of highly influential sports accounts in India. They apply the same methodology to curate a similar database for sportspersons in the USA

and do a comparative analysis of the engagements of sportspersons with politicians between the two countries. Similar methodology was used by Arya, Shora, and Pal (2021) to curate a dataset of influential business leaders in the USA, and did an analysis of their commentary on key issues related to Sustainable Development Goals on social media compared to influential business leaders in India, curated by snowballing the list in our database.

- Kommiya Mothilal et al. (2022) examine the Twitter engagement between Indian politicians and two sub-categories of influencers in our dataset - 'entertainers' and 'sportspersons'. They propose metrics to measure partisanship along different modes of engagement, analyze the discourse in engaged tweets, and study the public reception of such engagements. They find that the ruling party was more effective in reaching out to celebrities by shunning explicit partisan topics and subtly employing non-partisan narrative technique instead.

Data Description

The dataset we present has a main table with 11580 records. The table has 11 columns, the details of which are present in Table 3. Among them, our novel contribution are the fields of "category", "sub-category" and "type". Categories refer to a typology of the individuals based typically on their occupation. Further, to account for the presence of influencers from non-traditional backgrounds of celebrity, we refer to the "digital first" category. This refers to accounts that share little about their offline lives, and owe their popularity to their digital activity. For instance, the sub-categorisation of "fan accounts", "humour" (relating to meme accounts), and "informational" (e.g. an account on automated weather updates) bring forth the nature of the broad category.

Here, the "platform first" sub-category is of particular interest, as it contains numerous accounts whose offline lives are either unverifiable, or never mentioned in the first place. These accounts may often claim to hold certain professions, but almost exclusively chime in on political controversies or news on the platform. As studies of this dataset have indicated (Dash et al. 2021b,a), the continued involvement of platform celebrities in furthering partisan viewpoints raise further questions about grey area between layperson commentary and coordinated topic manipulation on Twitter.

In addition to the "digital first" category, the remaining broad fields we consider in the dataset have to do with the types of authority ascribed to commentators. Thus, "media" here refers to those involved in the creation and production of news, while "creatives" is a category for accounts engaged in arts, literature, film and TV production. The "civil society" category is distinct from the governmental and business realm and is ideally expected to to contribute to public discussions in an informative manner, relating to their specialist domains. Thus, lawyers, doctors, academics make up the category, apart from special interest groups and religious bodies. The last two categories of "sport" and "business" represent the fairly straightforward relationship these accounts have with the two occupations, either as organisations involved in the daily workings of the fields, or as individuals engaged in them.

Additionally, depending on the nature of studies that utilize this dataset, particular sub-categorisations can also be paired together for the unique requirements of research questions. For instance, a study on sporting commentary can include both the "sports" category, as well as journalists who contribute to public discussions about matches. Studies on policing can also refer specifically to the "law & policy", and "social worker" sub-categories to track the interaction of law enforcement accounts and advocacy campaigns online.

A detailed list of these categorical variables are presented in Table 2. Descriptions for each of these subcategories are presented in Table 1 (see below).

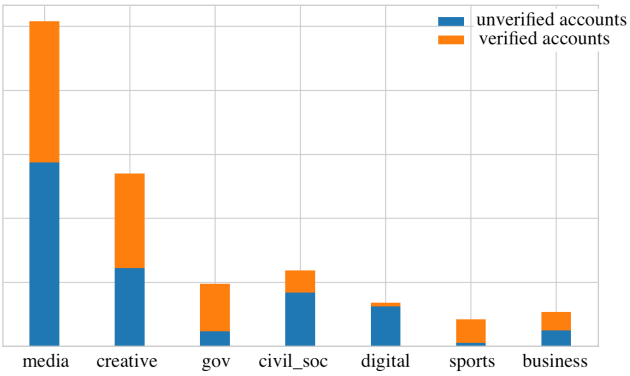

Figure 1: Category-Wise Counts

Analysis of Represented Accounts

We visualise all the Twitter users represented in out dataset across the 7 categories as separate category-wise scatterplots in Figure 2. To locate a user in a 2-dimensional plane, the x-axis (logarithmic) represents the number of accounts followed by a user and the y-axis (logarithmic) represents the number of followers of the user. The size of the point representing each user is proportional to the total number of tweets and retweets ever made by that user. All these public metrics were accessed via the Twitter API. We observe from the visualizations that:

- Users categorized as Civil Society tend to have narrower range of followers ($10^3 - 10^5$) than those categorized as Media or Creative ($10^3 - 10^7$)
- In contrast to the other categories, many Business accounts that typically tweet most actively tend to follow much fewer people ($< 10^2$)

Conclusion and Future Work

A novel database of highly networked individuals on Twitter is a window to better our understanding of narrative and influence on Social Media. With our focus on the Global South, we believe this dataset opens up new possibilities to understand interaction and dialogue in India, especially on how influencers in various spaces of public life intersect

Categories	Sub Categories	Description
civil society	academic	academic/research publications separate from newspapers and magazines, individuals in academia, or individuals engaged in fiction or non-fiction authorship.
	social worker	individuals who are distinguished by their work and service towards social causes
	law policy	non-academic specialists engaged with policy and law-making
	religious organisation	entities associated with religious and spiritual sects
	research organisation	entities associated with academic research - includes publishing houses and think tanks
	specialist	individuals who have highly specialised areas of work and impact, which aren't covered in the other sub-categories
government	defence	Accounts of defence personnel - veterans and other military professionals
	bureaucracy	Accounts of Indian bureaucrats, IAS, and state-level administrative officers. (E.g.: DMs, Ministry Secretaries etc.) (IFS not included)
	police	District, state, railway, traffic, police accounts, control rooms.
	official	Nationalised entities such as national banks, administrative departments, NITI Ayog, branches of defence etc. Also includes official accounts of people with positions in the government
business	brand	Major product brand, differentiated from a product holding company or corporation.
	businessperson	Individual owner or leader of a business concern.
media	journalist	news anchors, columnists and others at major media houses
	media house	entities and organisations involved in journalism, news, media and advertising
creative	entertainment	includes artists, actors, musicians, comedians, reality TV talent, show hosts
	writer	authors, primarily deriving fame outside academic circles
	designer	artists and designers of prominence
sports	team	official accounts of sports teams - includes cricket, football, IPL teams etc
	person	accounts of sports persons
digital first	fan account	accounts in appreciation of celebrities
	humour	satire, parody, meme pages and accounts
	informational	includes accounts sharing facts, information, historical accounts, job opening updates etc.
	platform first	influencers for whom the primary source of popularity is through an internet platform (e.g.: YouTubers, TikTokers, Twitter-famous civilians.)

Table 1: Subcategory Description

with vested interests such as politicians, or impact the public discourse, such as influencing the conversation on certain topics. Along the lines of previous work that have used subsets of this dataset, we are also working on analysing various other categories of users present in our dataset, such as government and defence.

This dataset and its sub-categories are also meant to be a living resource, since new influencers will get added, and categories will not only need to be updated for specific accounts, but the entire notion of a category may need to be rethought, reframed. What we also do here is provide a reasonably exhaustive, and closely vetted collection of seed-accounts which can be used to iteratively build a larger set by snow-balling through their immediate friend networks. For instance, if one were interested in dramatically increasing the "business" category, one could quickly snowball that into a much larger set to do a deeper study of business behavior on social media in India. This opens up the possibility for deeper dives into running domain-specific studies that need to characterize how an entire universe of users in a category behave on social media. We plan on using similar techniques to build upon specific sub-categories and study their interactions in greater detail.

Category	Counts	Subcategory	Counts
Media	5079	Journalist	4099
		Media House	980
Creative	2698	Entertainer	2622
		Writer	66
		Designer	10
Civil Society	1184	Specialist	732
		Law & Policy	160
		Social Worker	130
		Research Org	77
		Academic	69
		Religious Org	16
Government	977	Police	205
		Bureaucrat	223
		Organisation	405
		Defence	29
		Official	115
Digital First	684	Platform First	431
		Fan Account	91
		Informational	100
		Humor	62
Sports	423	Team	82
		Person	341
Business	535	Brand	146
		Person	389
Total	11580		11580

Table 2: Subcategory-Wise Counts

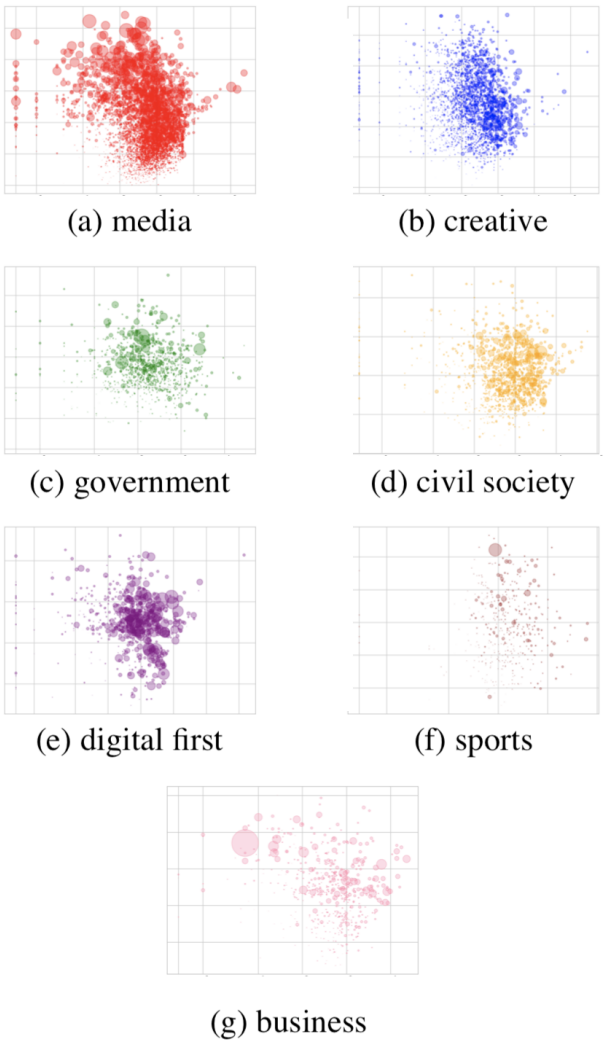

Figure 2: Visualising 'measures-of-influence' of represented users across all categories. The x-axis represents the number of friends, y-axis represents the follower count and the size of a bubble is proportional to the tweeting volume

Ethical Statement

Composition

The dataset contains 11580 records, each of which are of Twitter users. Additional details of its composition has been described in an earlier section and in Tables 2,3 and 1. By its nature, the dataset is a non-random sample from the set of all Twitter Users. Only 2 fields ('url' and 'location') may have missing information for some instances - these are self-reported values and their availability depends entirely on whether the Twitter user has shared any such value. All other fields have no missing values. By it's nature, the dataset can be used to identify individuals, more specifically, Twitter users. Apart from certain self-reported fields, such as location - no other field contains any sensitive information. All fields in the dataset are derived from the Twitter API and is public to general audiences.

Since the dataset involves manual annotation of the "category" and "subcategory" fields, a small margin of human-error is to be expected. In the absence of any real measure of ground-truth, we are unable to evaluate this error percentage quantitatively.

Uses

This dataset, subsets and expansions thereof, have been used, and are currently being used in several academic projects. We list some published research using this dataset in a prior section. We also outline other possible uses in the aforementioned section on applications. We note that all uses of the dataset must be cognizant of unavoidable errors that may have crept in as a result of manual annotation. Certain fields ("username","description","followers" etc.) may change over time – we recommend updating these fields using the Twitter API and the "id_str" field before the dataset is used in any application involving these fields. We reiterate that the primary contribution of this dataset is the annotation of "categories" and "subcategories".

Distribution and Maintenance

We have hosted the dataset on Dataverse. It may be accessed directly from Dataverse for any uses. Our main fields of interest are 'category' and 'subcategory' - these will not be updated unless to correct labelling errors. In case of the latter, kindly contact the authors. By it's nature, this dataset may be augmented and expanded upon by users to tailor it to their

Field Name	Description	Type	Unique Counts
id_str	Unique Twitter ID	Unique String	11580
created_at	Account creation date	UTC Datetime	11578
name	Twitter name	String	11442
username	Twitter Handle	Unique String	11580
description	Twitter bio	String	11124
followers	Number of followers	Numeric	10081
url	URL from profile text	String	7515
location	Location	String/Categorical	2408
type	Individual or Entity	Categorical	2
verified	Blue-ticked account	Boolean	2
category	Primary Industry	Categorical	7
sub_category	Primary Occupation	Categorical	24

Table 3: Dataset Columns

specific needs. The authors cannot guarantee or verify such modifications, however.

FAIRness of the Dataset

We host our dataset (along with metadata) on Dataverse. Dataverse is an open-source data repository software used widely, which provides a convenient way for dataset authors to adhere to FAIR (Wilkinson et al. 2016) principles. Our attempt at the same involves:

- **Findability**: Dataverse assigns a unique DOI (Document Object Identifier) when a dataset is published. This DOI resolves to a landing page with metadata, data files, terms, waivers and licenses.
- **Accessibility**: Dataverse provides public machine-accessible interfaces to search the data, access the metadata and download the data files, using a token to grant access when data files are restricted ('A').
- **Interoperability and Resuablility**: Dataverse offers the metadata at following 3 levels of hierarchy:
  1. data citation metadata (DataCite or Dublin Core)
  2. domain-specific metadata
  3. file-level metadata

Acknowledgements

We thank the editors and reviewers for their valuable feedback.

References

Al Zamal, F.; Liu, W.; and Ruths, D. 2021. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. *Proceedings of the International AAAI Conference on Web and Social Media* 6(1): 387–390. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14340.

Arya, A.; Shora, S. R.; and Pal, J. 2021. *Beyond Business: A Poster Contrasting CEO Activism on Social Media in India and the United States*, 432–436. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384537. URL https://doi.org/10.1145/3460112.3471983.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You Are Where You Tweet: A Content-Based Approach to Geo-Locating Twitter Users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, 759–768. New York, NY, USA: Association for Computing Machinery. ISBN 9781450300995. doi:10.1145/1871437.1871535. URL https://doi.org/10.1145/1871437.1871535.

Cohen, R.; and Ruths, D. 2021. Classifying Political Orientation on Twitter: It's Not Easy! *Proceedings of the International AAAI Conference on Web and Social Media* 7(1): 91–99. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14434.

Conover, M. D.; Goncalves, B.; Ratkiewicz, J.; Flammini, A.; and Menczer, F. 2011. Predicting the Political Alignment of Twitter Users. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 192–199. doi:10.1109/PASSAT/SocialCom.2011.34.

Dash, S.; Grover, R.; Shekhawat, G.; Kaur, S.; Mishra, D.; and Pal, J. 2021a. Insights Into Incitement: A Computational Perspective on Dangerous Speech on Twitter in India. *arXiv preprint arXiv:2111.03906* .

Dash, S.; Mishra, D.; Shekhawat, G.; and Pal, J. 2021b. Divided We Rule: Influencer Polarization on Twitter During Political Crises in India. *arXiv preprint arXiv:2105.08361* .

Dash, S.; Shekhawat, G.; Akbar, S. Z.; and Pal, J. 2021c. Extremism & Whataboutism: A Case Study on Bangalore Riots. *arXiv preprint arXiv:2109.10526* .

Kommiya Mothilal, R.; Mishra, D.; Nishal, S.; Lalani, F.; and Pal, J. 2022. Voting with the stars: Analyzing Partisan Engagement between Celebrities and Politicians in India. *Proceedings of the ACM on Human-Computer Interaction* .

Mahmud, J.; Nichols, J.; and Drews, C. 2021. Where Is This Tweet From? Inferring Home Locations of Twitter Users. *Proceedings of the International AAAI Conference on Web and Social Media* 6(1): 511–514. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14299.

Mishra, D.; Sen, R.; and Pal, J. 2021. Sporting the government: Twitter as a window into sportspersons' engagement with causes in India and USA. *CoRR* abs/2109.07409. URL https://arxiv.org/abs/2109.07409.

Panda, A.; Gonawela, A.; Acharyya, S.; Mishra, D.; Mohapatra, M.; Chandrasekaran, R.; and Pal, J. 2020. NivaDuck - A Scalable Pipeline to Build a Database of Political Twitter Handles for India and the United States. 200–209. doi:10.1145/3400806.3400830.

Pennacchiotti, M.; and Popescu, A.-M. 2021. A Machine Learning Approach to Twitter User Classification. *Proceedings of the International AAAI Conference on Web and Social Media* 5(1): 281–288. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14139.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; and Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. In *Scientific Data*. New York, NY, USA: Nature. doi:10.1038/sdata.2016.18. URL https://doi.org/10.1038/sdata.2016.18.