A Large-Scale Longitudinal Multimodal Dataset of State-Backed Information Operations on Twitter

Xiaobo Guo, Soroush Vosoughi

Department of Computer Science, Dartmouth College Hanover, New Hampshire
{xiaobo.guo.gr, soroush.vosoughi}@dartmouth.edu

Abstract

This paper proposes a large-scale and comprehensive dataset of 28 sub-datasets of state-backed tweets and accounts affiliated with 14 different countries, spanning more than 3 years (from 2015 to 2018), and a corresponding "negative" dataset of background tweets from the same time period and on similar topics. To our knowledge, this is the first dataset that contains both state-sponsored propaganda tweets and carefully collected corresponding negative tweet datasets for so many countries spanning such a long period of time.

Introduction

Propaganda is a form of communication that attempts to achieve the response that furthers the desired intent of the propagandist (Jowett and O'donnell 2018) by means of selectively presenting facts to encourage a particular synthesis or perception, or using objective language to evoke emotional rather than rational responses of the audience. One of the most harmful types of Internet propaganda is state-sponsored propaganda [1]. This is because considering the large resources at the disposal of the state-sponsored propagandists, countries can sway public opinions by flooding social media with their messages (Fisher 2020).

To detect state-sponsored propaganda on social media and minimize its harmful effects, much research has been conducted over the past several years such as analyzing the spread of propaganda (Zannettou et al. 2019b; Badawy et al. 2019), analyzing the user features of state-sponsored trolls (Zannettou et al. 2019a; Badawy, Lerman, and Ferrara 2019; Volkova et al. 2017) and identifying state-sponsored propaganda trolls (Luceri, Giordano, and Ferrara 2020; Miao, Last, and Litvak 2020; Orlov and Litvak 2018) or content (Guo and Vosoughi 2020) on social media.

While these works tackle various compelling research questions, they all require an annotated dataset of tweets or Twitter users as input. Unfortunately, high-quality annotated datasets of both positive (state-sponsored) and negative (not state-sponsored) data covering multiple countries are a rare commodity despite being essential for improving and reliably measuring model performance. Two datasets concerning state-sponsored propaganda on social media have been created and made publicly available to be used by researchers. These datasets, one comprised of Twitter data[2] and the other comprised Reddit data [3], cover various countries and topics. However both of them include only the state-sponsored accounts without a negative set of "normal" (not state-sponsored) accounts. The absence of this negative dataset means that these datasets can be used only for analyzing the difference between posts from different organizations, and not for identifying state-sponsored posts. To solve the problem, other works have built their own negative datasets at user-level or post-level.

Some have built baseline data at the user-level by filtering users with specific factors such as the distribution of the average number of tweets per day posted (Zannettou et al. 2019a) and the combination of language and location(Alhazbi 2020). Badawy et al.(Badawy, Lerman, and Ferrara 2019) used certain hashtags and keywords associated with each major presidential candidate of the 2016 US election to filter users. However, all these works ignore the topic preference of users, which might cause analysis or predictive modeling to be biased as there will be information leakage from the topical information. Other researchers (Broniatowski et al. 2020; Guo and Vosoughi 2020; Volkova et al. 2017; Chang et al. 2021; Orlov and Litvak 2018) overcame this problem by building the baseline data at the post-level to ensure that negative and positive tweets both contain similar hashtags and keywords. While these works partly solve the problem of topic distribution, they are only focused on limited topics or countries which limits the usage of these dataset.

Therefore we propose an original, large-scale and comprehensive dataset focusing on state-sponsored propaganda on Twitter. Compared with previous work, our dataset ensures greater robustness and generalizability of the models and analysis using it because of its less biased data sampling, topic similarity and the temporal alignment of the of positive and negative tweets, and the fact that it covers multiple coun-

[1] We use the terms state and country interchangeability in this paper.

[2] https://transparency.twitter.com/en/reports/information-operations.html

[3] https://www.reddit.com/r/announcements/comments/8bb85p/reddits_2017_transparency_report_and_suspect/

tries (14) and languages. The negative tweet dataset includes 67 different languages, with the top 3 most frequent languages being English (33.73%), Arabic (29.14%), and Spanish (6.94%). The positive tweet dataset includes 61 different languages, with the top 3 most frequent languages being Arabic (30.16%), Spanish (16.19%), and English (12.82%).

The main characteristics of this dataset are:

- 28 different sub-datasets covering 14 countries.
- IDs and metadata for 22,850 state-sponsored accounts and 667,803 "normal" accounts.
- IDs and metadata for 10,189,437 text-only positive (i.e., propaganda) tweets and 2,575,521 positive tweets with images.
- IDs and metadata for 1,144,614 text-only negative (i.e., non-propaganda) tweets and 202,732 negative tweets with images.

Data Access

Our dataset is hosted by Harvard Dataverse with the following link: https://doi.org/10.7910/DVN/NO3I34.

Data Collection

Our dataset is comprised of positive data which is the tweets (and corresponding account information) published by accounts affiliated with the state-sponsored organizations and negative data which is the tweets (and corresponding account information) published by background users. We leverage hashtags to filter tweets to ensure that positive and negative tweets are of similar topics.

State-Backed Data (Positive Set)

The positive data is collected from the Twitter Transparency report about information operation [4]. The original data from Twitter includes the tweet information, user information, and the country of origin. In our dataset, we treat the archives from the same countries of different time periods as different sub-datasets because they are identified by Twitter separately. Details of the 28 sub-datasets can be found in Table 1.

Background Data (Negative Set)

The background data is collected from the *Twitter Stream Grab of Internet Archive*[5], which is a simple collection of tweets grabbed from the general Twitter stream. The Twitter Stream Grab makes use of the stream API provided by Twitter to randomly sample 1% of real-time tweets. Since our work dataset is mainly about the difference in the content of state-sponsored propaganda, compared to background tweets, any tweets in the background data containing identical text or images to the state-sponsored propaganda is removed.

Filtering Tweets

For both positive and negative data, we only keep the tweets between March of 2015 and December of 2018 (except December 2017, for which there is no data from the Internet Archive to be used for the negative set). Tweets containing only URLs, the retweet mark('RT') or the mention mark('@') (and nothing else) are removed.

To ensure the topics of positive and negative tweets of each month of each sub-dataset are similar, on a monthly basis, we filter tweets by important hashtags. The importance of each hashtag of each tweet is one divided by the number of hashtags of that tweet. Then we add the importance of each hashtag of all tweets in one month to calculate the monthly importance of the hashtags and generate the most important 15 hashtags for that month (sometimes the number will be less than 15 due to lack of activity for that month). After extracting the 15 most important hashtags, we remove all tweets in that month from the positive and negative sets which do not contain any of these hashtags. This ensures that both the negative and positives sets are generally about the same topic (insofar as hashtags capture topics).

To ensure that negative tweets are different from the positive tweets, we rely on the text and images to filter the negative tweets. To compare textual content, we processed the text of tweets by replacing URLs in the text with 'URL', removing the retweet mark ('RT') and the mention information (mention mark, '@', and mentioned users). To compare the images, we apply dHash which focuses on the gradients of images method[6] to generate hash values for each image. For each negative tweet, if its processed text or the hash value of its images are same with any positive tweets, it will be removed. This ensures that the content of the negative and positive sets are not identical, which is important when for instance training a content-based propaganda detection model.

We also check the number of users in the negative set that have been suspended or deleted. The total number of users in the negative accounts is 667,803 of which 92,649 have been suspended by Twitter, and 75,125 have been deleted. Note that these users were not identified as state-sponsored propagandist accounts by Twitter and have been presumably removed for other reasons.

Data Exploration

In this section, we first introduce the structure and the scale of our dataset. Then, we conduct tweet-level and user-level analysis for preliminary exploration of differences between the positive and negative data.

Structure of User and Tweet Data

Our dataset includes tweets and users corresponding to the tweets. All data is in a JSON format and we present them here in table format for convenience.

Table 2 shows an example tweet from our dataset. Our dataset includes the tweet ID (tweet id), the user ID (user id),

[4] https://transparency.twitter.com/en/reports/information-operations.html

[5] https://archive.org/details/twitterstream

[6] https://pypi.org/project/ImageHash/

Dataset	Class	# accounts	# of text-only	# with images
Internet Research Agency (October 2018)	positive	2,017	837,707	3,167
	negative	41,540	55,988	11,165
Iran (October 2018)	positive	484	52,531	44,835
	negative	50,224	46,724	18,193
Bangladesh (January2019)	positive	7	981	174
	negative	4,797	4,387	612
Iran (January 2019)	positive	1,752	506,127	128,465
	negative	22,516	19,239	10,386
Russia (January 2019)	positive	172	151,834	32,019
	negative	66,723	61,450	25,208
Venezuela (January 2019 set 1)	positive	707	450,099	1,455,179
	negative	57,975	72,798	20,400
Venezuela (January 2019 set 2)	positive	396	29,358	19,561
	negative	24,681	18,652	7,954
United Arab Emirates (March 2019)	positive	2,296	146,454	47,911
	negative	31,728	27,935	7,771
Ecuador (April 2019)	positive	597	63,069	7,250
	negative	21,801	17,803	6,314
Saudi Arabia (April 2019)	positive	4	32	24
	negative	58	36	22
United Arab Emirates and Egypt (April 2019)	positive	90	66,673	2,575
	negative	31,254	27,918	9,000
Iran (June 2019 set 1)	positive	121	124,120	40,017
	negative	36,284	33,471	14,151
Iran (June 2019 set 2)	positive	164	132,870	38,722
	negative	20,295	16,761	6,766
Iran (June 2019 set 3)	positive	1,454	23,566	6,315
	negative	29,226	23,260	8,731
Venezuela (June 2019)	positive	32	1,285	5,680
	negative	14,306	9,968	4,937
China (July 2019 set 1)	positive	469	273,602	9,055
	negative	63,478	46,053	31,836
China (July 2019 set 2)	positive	49	47,363	837
	negative	33,429	33,504	10,568
China (July 2019 set 3)	positive	455	514,125	46,905
	negative	162,060	151,277	70,525
Saudi Arabia (October 2019)	positive	2,412	753,514	98,616
	negative	88,465	78,972	35,361
Egypt (February 2020)	positive	829	891,204	138,352
	negative	33,682	28,263	10,407
Honduras (February 2020)	positive	738	133,405	10,635
	negative	17,737	12,075	6,738
Indonesia (February 2020)	positive	272	288,071	16,501
	negative	81,771	69,159	34,966
SA_EG_AE (February 2020)	positive	439	1,804,106	23,579
	negative	60,302	54,412	23,777
Serbia (February 2020)	positive	3,280	1,907,595	276,177
	negative	8,749	5,923	3,834
Ghana and Nigeria (March 2020)	positive	5	412	48
	negative	1,363	1,072	329
China (May 2020)	positive	9	107	20
	negative	615	511	120
Turkey (May 2020)	positive	2,946	859,212	84,673
	negative	22,976	16,734	8,564
Russia (May 2020)	positive	654	130,015	38,229
	negative	17,122	13,832	6,845

Table 1: Detail of the dataset, including the number of accounts (# of accounts), number of text-only tweets (# of text-only tweets), and the number of tweets with images (# of tweets with images) for the positive and negative data of each sub-dataset. The dates of the datasets corresponds to when they were released. SA_EG_AE (February 2020) includes tweets from three middle eastern countries (Saudi Arabia, Egypt and United Arab Emirates) combined by Twitter.

Title	tweet id	user id	subdataset	class	name	tweet time	account lang	tweet lang
Value	xx	xx	China (May 2020)	positive	xx	2018-04-16 03:59	zh-cn	en
Title	# of likes	# of retweets	hashtags	urls	mentions	images	image hashes	length
Value	0	0	[VersaceTribute]	[]	[xx, xx]	[http://xx]	[xx]	19

Table 2: An example tweet item. Note that we divided the row into two columns and replaced true tweet id, user ids, name, images and hashes of images with ‘xx’ since they are long.

Title	id	screen name	subdataset	class	location	creation date	lang
Value	xx	xx	Russia (May 2020)	negative	[]	2009-12-21	[ru]
Title	# min followers	# max followers	# min friends	# max friends	profile	age (days)	
Value	732	788	7	8	[#Troubleshooter]	2558	

Table 3: An example user item. Note that we divided the row into two columns and replaced true id, screen name with ‘xx’ since they are long.

the sub-dataset the tweet belongs to (sub-dataset), the class of data (class), the display name of the user (name), the time the tweet is published (tweet time), the preferred language of the user at that time (account language), the language of the tweet (tweet lang), the number of quotes likes (# of likes), the number of retweets (# of retweets), the list of hashtags (hashtags), the list of mentioned users (mentions), the list of URLs (urls), the list of filenames of images (images), and the list of hashes of images (image hashes). For positive tweets, the user id and the display name is hashed by Twitter.

In Table 3, we show an example user from our dataset. Our dataset includes user id (id), screen name (name), sub-dataset the user belongs to (sub-dataset), the class of user (class), the list of reported locations (location), creation date of the user (creation date), the list of languages the user prefers (lang), the range of the number of followers during the time period (# of min followers and # of max followers), the range of the number of friends during the time period (# of min friends and # of max friends), the list of profile descriptions overt time (profile) and the age of users calculated backwards from Dec 31st 2018 (age (days)). The ids for the negative users are the actual ones, while those of positive users are hashed by Twitter. For location, language, and profile description data, list format is used to represent all data in the time period. Because the number of followers and friends may vary during the time period, we use ’min’ and ’max’ to represent the minimum number and maximum number of followers and friends.

Tweet-Level and User-Level Analysis

As a preliminary analysis of the importance of tweet-level features in distinguishing state-backed and background tweets, we conduct a logistic regression analysis on all of our data. The dependent variable for our analysis is a binary indicator of whether a tweet is state-backed or background. The independent variables in our model are the following tweet features:
“subdataset”, “# mention”, “# hashtags”, “Length”, “Has Image”, “Has URL”, and “Language same” (whether the “tweet lang” and “account lang” are same). Note that since “subdataset” is categorical, it is represented by 28 binary dummy variables. We do not include “# of likes” and “# of retweets” since they cause “singular matrix” error as they are 0 for the vast majority of tweets.

Table 4 shows the effect (coefficient) and the statistical significance of each of the features in our logistic regression. The pseudo R-squared of the logistic regression is 0.2151. We can observe that all these features are to some level predictive of whether a tweet is state-sponsored or not (as can be seen in Table 4, all the features are statistically significant). Note that this is a preliminary analysis which does not take into account possible confounding variables. More detailed analysis is needed to better understand the effect of each feature. For brevity, we do not show the 28 “subdataset” dummy variables in the table.

The same analysis on user-level differences between state-sponsored and background accounts showed the user account age to be the only significant feature.

The age of a user account is calculated as the number of days between the date the account was created and Dec 31st 2018, which is the end date of our dataset. For most sub-datasets, we observe that the average age of users in the negative dataset is older than that of users in the positive dataset, except for China (July 2019 set 2), China (July 2019 set 3), Ecuador (April 2019), and Russia (May 2020). We also observe that for most sub-datasets, the average age of users in the positive dataset is less than four years. Taken together, these two phenomena demonstrate that most user account in the positive dataset are relatively young and seem to be created for a particular purpose (i.e., pushing a certain agenda). In Figure 1, we show the distribution of account ages for both the positive and negative accounts. We can observe that the age of positive accounts is on average lower than that of negative accounts. This makes intuitive sense as accounts in the positive dataset are typically created as needed for the purpose of spreading propaganda.

Limitations

Data Bias

Our dataset is an excellent resource to study state-sponsored propaganda on Twitter due to its large scale, its topic variety, its multi-modality, and the fact that it encompasses tweets from multiple countries and languages. However, because our dataset focuses on certain topics (i.e., hashtags), the be-

	coef	std err	z	P>$|z|$	[95.0% Conf Int.]	
# mentions	-0.2840	0.001	-311.346	0.000	-0.286	-0.282
# hashtags	-0.1063	0.000	-221.382	0.000	-0.107	-0.105
Length	0.0074	0.000	40.612	0.000	0.007	0.008
Has Image	-0.3688	0.003	-145.914	0.000	-0.374	-0.365
Has URL	-0.8331	0.003	-319.539	0.000	-0.838	-0.828
Language same	-0.4807	0.003	-188.539	0.000	-0.486	-0.476

Table 4: Results of the logistic regression model estimating whether a tweet is state-sponsored as a function of variables shown in the first column. The pseudo R-squared of the model is 0.2151 .

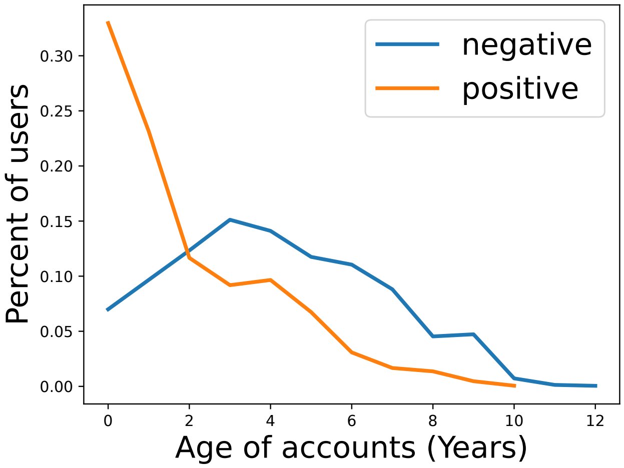

Figure 1: The distribution of age for positive and negative accounts.

havior and personal characteristics represented in the chosen tweets might not represent users comprehensively. Therefore, to conduct analysis at user-level, researchers should collect more tweets from the user timelines to reduce this potential bias. For users in the positive datasets, the full data can be collected from the archive provided by Twitter[7], and for users in the negative datasets, the timeline can be collected by the Twitter API based on the ID of users provided.

Additionally, due to our sampling process, our dataset should not be applied to any domains other than studying state-backed tweets. This is because of the following potential biases: First, all users in the positive datasets are state-sponsored users and do not represent the average Twitter user. Second, all users in the negative dataset tweet about certain topics which might not be representative of Twitter users as a whole. Third, while only around 10% of normal tweets are related to political topics (Colleoni, Rozza, and Arvidsson 2014), most of the tweets in our dataset are political. Depending on the need of researchers, they may benefit from different sampling strategies. In these cases, we recommend the researchers use the publicly available data to create samples that best fit their needs.

Primary Data Errors

Since none of our data is directly collected from Twitter via their API, the errors in the primary (i.e., raw) data will unfortunately be extremely hard to correct because the original positive tweets are no longer available for public access, and are only available through the archives provided by Twitter.

The most common error encountered in our primary data is the existence of negative values for certain numeric features that cannot be negative. For example, the # of retweets should always be larger than 0, but we observe that the # of retweets for some tweets is erroneously set to be smaller than 0.

Recommendations for Usage

Keeping in mind the aforementioned analyses and limitations, below we propose two key areas which could be investigated using our dataset.

Identification of State-Sponsored Propaganda

Detecting state-sponsored propaganda on social media is an important and timely topic of research that has the potential to have great impact. Training robust and generalizable machine learning models for detection of state-sponsored propaganda requires large labeled datasets spanning different time periods, topics, and organizations. Our dataset is ideal for training such models as it satisfied these requirements.

Moreover, the task of identifying the source (i.e., the country) behind state-sponsored propaganda is also an interesting research problem. Our dataset simplifies training multi-class prediction models by providing labeled data from different countries.

Analysis of State-Sponsored Propaganda

Apart from identification, analyzing state-sponsored propaganda is of great importance. Analysis of this data can strengthen our understanding of the mechanisms used by various states to propagate propaganda on social media. This can help devise policies to dampen the influence of such propaganda.

Since our dataset includes user and content information about propaganda campaigns (and corresponding background data), it is ideal for analysis of the techniques used by different state-backed propaganda agencies. Our dataset for instance can be used to study different tactics used by different organizations when discussing similar topics and

[7] https://transparency.twitter.com/en/reports/information-operations.html

opinions. Additionally, our dataset can be used to study issues that are being supported or attacked by certain countries to shed light on the geopolitical stances and goals of these countries.

Finally, given the temporal, topical, and organizational diversity of our dataset, it can be used to gain a deeper understanding of propaganda messaging, accounting for organizational, topical and temporal features.

Conclusion

In this work, we introduced a large-scale, longitudinal, and diverse dataset for studying state-backed information operations on Twitter, covering 14 countries and more than 3 years (2015-2018). Our dataset includes both content and user information and is unique in that the positive propaganda dataset has been matched with a topically and temporally aligned negative background dataset.

We conducted preliminary analysis of tweet-level and user-level features of the state-sponsored and background accounts. The results showed that there are statistically significant differences between the two types of data with respect to some of the features. These results could be useful for identifying and analyzing state-sponsored propaganda in future work.

Ethics Statement

The positive dataset presented in this paper is from Twitter's transparency report of information operations, and the negative dataset is from the Internet Archive, both of which are publicly available with strict privacy standards. We do not release any data that is not part of these two public repositories.

The main ethical consideration related to the use of our dataset is the problem of false positives, which refers to wrongly labeling an "innocent" user or tweet as being state-sponsored. Considering the potential real-world negative impact of false positive prediction for users (ranging from social stigma to putting their lives in danger) and society as a whole (e.g., harming freedom of speech and the diversity of ideas), it is important to be cautious when creating models for identifying state-sponsored accounts and propaganda. Furthermore, according to our definition of state-sponsored tweets, any tweet that is shared by state-sponsored accounts is labeled as being state-sponsored, even if the original tweet that was shared came from a non-state-sponsored account. Thus, it is extremely important for any paper relying on our dataset to include thorough analysis of false positives and explicitly discuss issues mentioned here.

References

Alhazbi, S. 2020. Behavior-Based Machine Learning Approaches to Identify State-Sponsored Trolls on Twitter. *IEEE Access*, 8: 195132–195141.

Badawy, A.; Addawood, A.; Lerman, K.; and Ferrara, E. 2019. Characterizing the 2016 Russian IRA influence campaign. *Social Network Analysis and Mining*, 9(1): 31.

Badawy, A.; Lerman, K.; and Ferrara, E. 2019. Who falls for online political manipulation? In *Companion Proceedings of WWW 2019*.

Broniatowski, D. A.; Kerchner, D.; Farooq, F.; Huang, X.; Jamison, A. M.; Dredze, M.; and Quinn, S. C. 2020. The covid-19 social media infodemic reflects uncertainty and state-sponsored propaganda. *arXiv preprint arXiv:2007.09682*.

Chang, R.-C.; Lai, C.-M.; Chang, K.-L.; and Lin, C.-H. 2021. Dataset of Propaganda Techniques of the State-Sponsored Information Operation of the People's Republic of China. *arXiv preprint arXiv:2106.07544*.

Colleoni, E.; Rozza, A.; and Arvidsson, A. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication*, 64(2): 317–332.

Fisher, A. 2020. Demonizing the enemy: the influence of Russian state-sponsored media on American audiences. *Post-Soviet Affairs*, 36(4): 281–296.

Guo, X.; and Vosoughi, S. 2020. Multi-modal Identification of State-Sponsored Propaganda on Social Media. *arXiv preprint arXiv:2012.13042*.

Jowett, G. S.; and O'donnell, V. 2018. *Propaganda & persuasion*. Sage Publications.

Luceri, L.; Giordano, S.; and Ferrara, E. 2020. Don't Feed the Troll: Detecting Troll Behavior via Inverse Reinforcement Learning. *arXiv preprint arXiv:2001.10570*.

Miao, L.; Last, M.; and Litvak, M. 2020. Detecting Troll Tweets in a Bilingual Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 6247–6254.

Orlov, M.; and Litvak, M. 2018. Using behavior and text analysis to detect propagandists and misinformers on twitter. In *Annual International Symposium on Information Management and Big Data*, 67–74. Springer.

Volkova, S.; Shaffer, K.; Jang, J. Y.; and Hodas, N. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 647–653.

Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019a. Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web. In *Companion proceedings of the 2019 world wide web conference*, 218–226.

Zannettou, S.; Caulfield, T.; Setzer, W.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019b. Who let the trolls out? towards understanding state-sponsored trolls. In *Proceedings of the 10th acm conference on web science*, 353–362.