Datasheet for "Coordinated Reply Attacks in Influence Operations: Characterization and Detection"

Manita Pote

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The 'Reply Coordination Detection' dataset was curated to study the coordinated replies in the context of influence operation on Twitter and to build a machine learning model to identify tweets that get malicious coordinated replies as well as the repliers involved in such tasks.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by the first author of the paper on behalf of the Observatory in Social Media at Indiana University, Bloomington.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

This work was supported in part by Knight Foundation, Craig Newmark Philanthropies, and DARPA (contract HR001121C0169); and by Lilly Endowment, Inc. through funding for the Indiana University Pervasive Technology Institute.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

There are multiple data files related to each research question. Therefore, we present the details on each of these data files.

RQ1_target_annotation.csv

Each record in this file represents the manual annotation of the profession and country of a user or target for the Serbia and Egypt campaigns. This file has following columns:

- userid
- profession_label

- country_label
- campaign

Values of the 'campaign' column are either Serbia or Egypt. There are 1,547 records.

RQ1_target_non_target_tweets.csv

Each record in this file represents an English translated targeted or non-targeted tweet in the Serbia and Egypt campaigns. This file has the following columns:

- tweetid
- english_text
- type
- campaign

The 'type' column has two values: (1) *target*, which represents a targeted tweet, or (2) *non_target*, which represents a non-targeted tweet. The 'campaign' column has two values, Serbia or Egypt. There are 36,885 records.

RQ1_target_follower_following_count.csv

Each record in this file represents a target user and reports the numbers of followers and following of that target. There are 15,016 records. This file has the following columns:

- userid
- followers_count
- following_count

RQ1_number_of_reply_per_tweet.csv

Each record in this dataset represents a targeted tweet and reports the number of IO replies received by that tweet. There are 53,931 records. This file has the following columns:

- poster_tweetid
- reply_count

RQ1_num_targeted_tweet_by_IO.csv

Each record in this dataset represents a pair of users: a target (poster) and an IO replier. The record reports the number of tweets by the given poster that are targeted by the given IO replier. There are 148,450 records. This file has the following columns:

- poster_userid
- replier_userid
- count

RQ1_time_difference_of_reply.csv

Each record in this file represents a reply and its delay (time since tweet targeted by the reply). There are 148,450 records. This file has the following columns:

- replier_tweetid
- diff_min

RQ2_engagement.csv

Each record in this file represents a targeted or control tweet and reports the engagement received by that tweet. There are 7,732 records. This file has the following columns:

- tweetid
- retweet_count
- like_count

- quote_count
- type

The 'type' columns has two values : (1) *target* or (2) *control.*

RQ2_tweet_classifier_features.csv

Each record in this file represents a targeted or control tweet and reports features associated with that tweet. Please refer to the RQ2 section of the manuscript for the names of the features and how they are calculated. There are 7,732 records in this file. The 'tweet_label' columns indicates whether the tweet is targeted (1) or control (0) and is used to train/test the tweet classifier.

poster_tweetid_campaign_type.csv

Each record in this file represents a pair consisting of a tweet by a target and a reply to that tweet. The file reports the target user, the replier, which campaign the replier belongs to, a label for the type of replier, and a label for the type of original tweet. This file has 2,673,091 records with the following columns:

- poster_tweetid
- campaign
- replier_userid
- replier_label
- replier_tweetid
- type

The 'type' column indicates if the original tweet is targeted or part of the control. The 'replier_label' column indicates if the replier is a normal replier (0) or an IO replier (1).

RQ3_replier_classifier_features.csv

Each record in this file represents a replier and lists features used to classify that replier. The 'replier_label' column indicates if the replier is a normal replier (0) or an IO replier (1). The file has 881,918 records. For detail about the features, please refer to the RQ3 section of the paper.

RQ3_replier_info.csv

Each record in this file represents profile meta-data for a replier. The file reports the aggregated count of the replier's activities, their follower and following counts, age, and whether the replier is an IO replier (1) or a normal replier (0). There are 881,918 records. This file has the following columns:

- replier_userid
- activity_count
- replier_label
- following_count
- followers_count
- age

pos_cosine_with_replier_info.pkl.gz

Each record in this file represents a pair of replies to a tweet. The file reports the cosine similarity between the two replies, the original tweet that received the two replies, the two users who posted the replies, and their labels: 0 for normal repliers and 1 for IO repliers. This file has 1,260,848,920 records with the following columns:

- replier_label_x
- replier_label_y
- replier_userid_x

- replier_userid_y
- replier_tweetid_x
- replier_tweetid_y
- poster_tweetid
- cosine

How many instances are there in total (of each type, if appropriate)?

The numbers of records in each file are mentioned in the previous section.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

For RQ1, the dataset contains records for only two campaigns selected by the authors. For RQ2 and RQ3, the control data contains records about replies to tweets by target users limited to specific time intervals (see paper for details). With these exceptions, the dataset contains all records related to the selected influence operations reported by Twitter (see paper for details).

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The 'RQ2_tweet_classifier_features.csv' and 'RQ3_replier_classifier_features.csv' files contain features for the tweet and replier classifiers. For the full list of the features and information on how they are constructed, please refer to sections RQ2 and RQ3 in the paper. All other files contain derived data that are required to replicate the results of the paper.

Is there a label or target associated with each instance? If so, please provide a description.

Yes. Please refer to the individual file descriptions above.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

All data is preprocessed. Most raw data from the tweets is excluded. Only information needed to reproduce the results is included.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Yes. Records can be linked through user IDs and tweet IDs.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

We recommend 10-fold cross validation. Please refer to the details of the training setup in the paper.

Are there any errors, sources of noise, or redundancies in the

dataset? If so, please provide a description.

Not to our knowledge.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained in terms of reproducing the results. The original data for IO replies to targeted tweets comes from the Twitter Influence Operation dataset, which was publicly available until summer 2024 (https://web.archive.org/web/20240829231920/https://transparency.x.com/en/reports/moderation-research). The replies from normal repliers as well as control tweets were collected by the first author using the Twitter API. However, due to recent changes in Twitter/X's data sharing policy, the future availability of this data is uncertain. Currently, obtaining raw tweet and user metadata from Twitter/X using the IDs in our dataset is prohibitively expensive.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

The dataset contains tweet IDs and user IDs for IO users/tweets that have been suspended, and for control users/tweets that are public. This is consistent with Twitter/X terms. The dataset also includes annotations of a sample of target users based on their public profiles. None of this data should include any confidential information.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, some of the files with user IDs are related to people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

In the 'RQ1_target_annotation.csv' file, target Twitter accounts are identified by profession and country of origin. The distributions are reported in the paper.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Yes. We provide user IDs of Twitter accounts, which are public information. One can identify the individuals by querying the Twitter API.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No. The dataset contains derived data and Twitter account/tweet IDs are public information.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The replies from Influence Operation (IO) accounts were derived from Twitter's Influence Operation dataset, which was publicly available on the website of the Twitter Moderation Research Consortium (https://web.archive.org/web/20240829231920/https://transparency.x.com/en/reports/moderation-research) until summer 2024. All other data, including target profiles, normal replies, and control tweets, were collected using the Twitter Academic API. The data has been thoroughly validated to remove duplicates and ensure consistency in data types. Data about suspended accounts and inconsistent records (e.g., profiles with creation date before the creation of Twitter) have been removed to maintain data quality.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The target accounts were manually annotated by one of the authors. The annotation process is described in the RQ1 section of the paper. Other data were collected from the Twitter API using client software developed by the author. Extensive exploratory data analysis was conducted to gain insights into the data, with key findings summarized in the paper.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The targeted tweet dataset consists of a sample of tweets from Twitter Influence Operation (IO) campaigns, specifically those that have five or more replies from IO accounts. In addition, we collected all replies (without sampling) from normal users responding to the same targeted tweets. For the control dataset, we collected all replies (not sampled) to selected tweets by targeted users that had five or more replies from any users.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The first author collected the data.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data from IO accounts comes from an archived dataset. The normal replies to the targeted tweets and the control dataset were collected in the spring of 2023 but created over a period of years prior to that.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

The data collection was deemed exempt from review by the Indiana University IRB (protocols 12410 and 1102004860).

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, RQ1_target_annotation.csv, RQ3_replier_info.csv and RQ1_target_follower_following_count.csv are associated with Twitter profiles.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data were collected through the Twitter API (https://developer.x.com/en/docs/x-api).

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No, the individuals were not notified.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

No consent was required.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

n/a

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

The Indiana University IRB determined that the data collection consisted in observation of public behavior with minimal risks to subjects.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Our data processing pipeline includes cleaning (matching accounts in target and control datasets, balancing the classification dataset, etc.); removal of duplicate, suspended, and inconsistent accounts; missing data handling; and standardization of features. For the case studies, the tweets were translated into English from Serbian and Arabic.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is saved along with the preprocessed one for future consistency checks and reproducibility. It may be shared upon request if required for reproducibility and if this can be done in compliance with ethical/IRB policies and platform terms.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No. Software is available to reproduce the results in the paper starting from the cleaned data in our data repository.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Part of our data is derived from Twitter's Influence Operation dataset, which has been analyzed in other studies.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

We are not aware of all the papers that used the Twitter's Influence Operation datasets. Papers from our lab are available at https://osome.iu.edu/research/publications

What (other) tasks could the dataset be used for?

The dataset could be used to develop more advanced models for detecting coordinated inauthentic behaviors.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The target dataset includes only tweets with five or more replies from IO accounts. We are not aware of any future uses that might result in undesirable harms.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Not to our knowledge.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, we are making the dataset publicly available for replication purposes.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

This datasheet is in the Zenodo repository that contains the dataset. The code is available in a Github repository (https://github.com/osome-iu/io-coordinated-replies), which points to the Zenodo DOI.

When will the dataset be distributed?

The dataset will be distributed after the publication of our manuscript. If you are reading this datasheet it means the dataset has been distributed.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Creative Commons Attribution 4.0 International (https://creativecommons.org/licenses/by/4.0/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

Twitter IO datasets were available under Twitter terms and policies.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors will be supporting and maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Contact author's homepage: https://manitapote.github.io/

Is there an erratum? If so, please provide a link or other access point.

n/a

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

If any errors are found in the future, the dataset will be updated by the author.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

The Indiana University IRB determined that the data collection consisted in observation of public behavior with minimal risks to subjects, therefore we are not aware of applicable limits.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

If the dataset is updated in the future, older versions will continue to be available on Zenodo.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Yes, the CC BY 4.0 license allows for adaptations with attribution. We may or may not be asked or be available to validate/verify derived datasets.