Curating the Twitter Election Integrity Datasets for Better Online Troll Characterization

Anonymous Author(s)
Affiliation
Address
email

Abstract

In modern days, social media platforms provide accessible channels for the interaction and *immediate* reflection of the most important events happening around the world. In this paper, we, firstly, present a *curated* set of datasets whose origin stem from the **Twitter's Information Operations**[1] efforts. More notably, these accounts, which have been already suspended, provide a notion of how state-backed *human trolls* operate.

Secondly, we present detailed analyses of how these behaviours vary over time, and motivate its use and abstraction in the context of deep representation learning: for instance, to learn and, potentially track, troll behaviour. We present baselines for such tasks and highlight the differences there may exist within the literature. Finally, we utilize the representations learned for behaviour prediction to classify trolls from *"real"* users, using a sample of non-suspended active accounts.

1 Background

The risks of political polarization have been a recurring theme in recent work, as a byproduct of the existence of malicious actors in social media. For instance, *echo chambers* form to create niches that amplify nuanced information [1]. Hence, detecting fake accounts is crucial to avoid these scenarios to develop: a recent approach utilizes community detection in their main basis [5]. These efforts stem from rather classical approaches that ensemble a plethora of models and try to identify the most important features that account for *human-bot* classification [7]. On a more applicable manner, recent work on the matter has served to spot and raise the awareness of a *"infodemic"* that comes along the COVID-19 pandemic [4, 3, 6].

Previous work on the *Twitter Election Integrity* (TEI) dataset has been reported recently. In [11] the authors analyze *10M posts* identified as Russian and Iranian state-sponsored trolls. Furthermore, they present a cross-platform *influence* model that quantifies, for instance, how likely is that events in a Twitter community influence subsequent ones within a Reddit community. in [10] a comparison is presented between users identified to have ties with the Russian *Internet Research Agency* and a random set of Twitter users; the authors find differences in terms of the content each group disseminate.

Closely related to the current work in [8] a troll classification task is presented over a dataset collected from the Internet Research Agency (IRA), targeting US-related events. The authors leverage temporal point processes within a mixture density network to capture characteristics from the users' behaviours. Finally, in [9], the authors analyze 1.8M images from Russian trolls in the the dataset to conclude

[1]Twitter's transparency website (`https://transparency.twitter.com/en/reports/information-operations.html`) serves as the main source for every release's information and description. Most notably, every hashed archive can be easily accessed via the same website.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

their posting activity matches with that of real-world events. They further provide claims on how state-sponsored trolls manage their image posts towards a specific target.

2 Dataset

The Twitter *Information Operations* database has been consistently renewed since late 2018. In line with their *transparency* objectives, and with the intention of helping the community to fight against *state-backed entities*, the aforementioned social platform invites members of governments and academia to further investigate, learn, and build technology using their archives. In October 2018, a set of $4,383$ accounts were made public to kick-start the program[2].

All released users have already being suspended. Moreover, all releases include both, a list of users and their metadata accompanied with a list of tweets, also with metadata such as the number of likes and retweets receives, along with the list of mentioned users and hashtags. While the main reasons for this data collection process could be summarized in rather *political* terms, it is the nature of each release itself what makes it challenging to directly exploit any state-of-the-art model on it. Most accounts have not really being automated as *bots*, hence this is an ubiquitous trace of activity processed directly by humans.

2.1 Collection Process

In order to work with the Twitter Election Integrity (TEI) data we have built a set of scripts that download and handle their preprocessing[3]. The counterpart of these trolls are the `user_mentions` they employ, that is, their 1-hop neighborhood. We make use of Twitter's *Academic* API[4] to perform any request as obtaining a significant amount of activity results a nontrivial effort.

Table 2 summarizes the total number of users and hashtags involved in the obtained data. The number of senders correspond to trolls reported originally inside TEI; on the other hand, 1-hop senders include active accounts which, for the purposes of this project, we take as a *real user* sample counterpart. The number of receivers combines hashtags and user mentions, while the number of tweets also considers duplicated uses of the aforementioned Twitter features to give the exact activity count. We focus our efforts only on sub-datasets that originated from *Russia*, the *Internet Research Agency* (IRA), and *China*.

	#senders	#receivers	#tweets	#hashtags	#user mentions
Russian	168.234	1850.009	2048.630	590.483	1259.526
Russian-1-hop	78.050	1602.487	1609.347	409.220	1193.267
IRA	181.118	6703.894	7070.404	0.003	6703.891
IRA-1-hop	59.604	1584.001	1775.280	507.192	1076.809
Chinese	233.120	2700.590	3695.759	1271.163	1429.428
Chinese-1-hop	46.634	2075.923	2105.192	633.273	1442.650

Table 1: Average number of *nodes* (senders, receivers), *links* (hashtags, user mentions), and *total activity* (tweets) of the TEI dataset, per five days.

3 Methdology

To construct a graph able to be processed by the subsequent models, we distinguish the set of **senders** $\mathcal{S}$ (users that emit a tweet) from the set of **receivers** $\mathcal{R}$ (either users that are mentioned or any

[2]Twitter's transparency website (`https://transparency.twitter.com/en/reports/information-operations.html`) serves as the main source for every release's information and description. Most notably, every hashed archive can be easily accessed via the same website.

[3]A *Google API Token* is needed to run and download the data.

[4]`https://developer.twitter.com/en/products/twitter-api/academic-research`

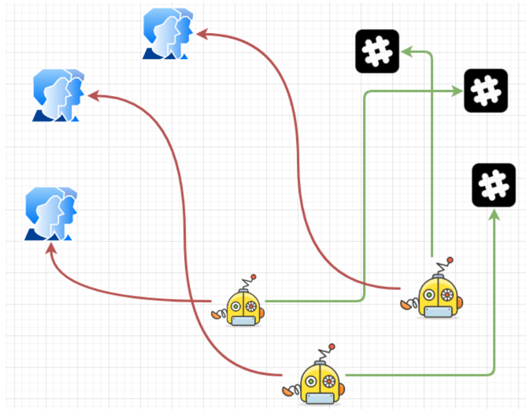

Figure 1: Sample figure caption.

hashtags). We follow a procedure that may include sampling the number of receivers or links, if any such quantity is out of balance with the number of senders; Figure 1 depicts the way the different types of nodes are connected; this is achieved, in summary, by the following process:

1. We fix a time interval $\delta = (t_{min}.t_{max})$ from which we extract all the tweets that were created no earlier than t_{min} and no later than t_{max}.
2. We examine the number of receiver users that result from the previous step.
    - If $|\mathcal{R}|$ is *significantly* greater than $|\mathcal{S}|$, we opt to randomly take a sample of mentioned users and hashtags, whose number *roughly* matches that of the senders.
    - We, thus, *"down-sample"* our chosen activity to balance the three types of nodes we are working with. This process helps to avoid biased predictions on certain classes.
3. We examine the number of existing links between senders and receivers.
    - If the number of links, regardless of repeated mentions, exceeds a limit parameter ℓ_E, we randomly select a subset of links.
    - Once again, this process helps us to control any undesired learned correlation on the final predictions.
4. Finally our (directed) adjacency matrix $A_{\mathcal{D}}$ indicates whether a sender account mentions a receiver account and whether it uses a certain hashtag.

For the link prediction pipeline, we need to construct node attributes beforehand. To leverage the heterogeneous nature of our proposed construction, where multiple types of nodes interact within each other, we utilize the `metapath2vec` algorithm [2], which biases random walks according to predefined node paths. For the purposes of this project, we identify four types of links defined by their incident nodes: `troll-uses-hashtag`, `troll-mentions-user`, `real-uses-hashtag`, and `real-mentions-user`.

We then use the **SEAL** (**S**ubgraphs, **E**mbeddings, and **A**ttributes for Link Prediction) [12] framework for link prediction on the aforementioned types of activities. Internally, a *node labeling* algorithm captures each node's role within its k-hop neighborhood. Moreover, we use a `min-pooling` layer to accumulate the learned features into node attributes, to later pass on a `multi-layer perceptron` that is trained to classify trolls from their 1-hop neighbours.

4 Experiments

We repeat a set of experiments by altering the length of the designated interval to construct a graph of interacting trolls and real users, as explained on Section 3. Table 2 summarizes our results, evaluated using F1 and accuracy scores. In this case we averaged over 5, 10, and 30 day intervals; moreover, link prediction scores seem better than those for node classification.

	F1/NC	accuracy/NC	F1/LP	accuracy/LP
Russian	0.73 ± 0.10	0.68 ± 0.06	0.78 ± 0.05	0.77 ± 0.04
IRA	0.64 ± 0.22	0.77 ± 0.07	0.85 ± 0.05	0.84 ± 0.05
Chinese	0.85 ± 0.07	0.75 ± 0.08	0.9 ± 0.04	0.85 ± 0.05

Table 2: Performance scores for the node classification (NC) and link prediction (LP) task, listed by dataset and by place of origin. We report F1-scores and accuracies averaged over every repeated experiment, defined by a sliding window over time that extracts the data in the way it is described previously.

5 Conclusion

In this project, we have taken a *structural* approach – within the jargon of graph representation learning – to train and learn some of the ubiquitous type of activities that fake users, namely *trolls* perform online. The importance of this task is justified by the recent reports of massive state-backed coordinated activities which target important political events, among other massive opinion changes. We were able to learn a state-of-the-art deep neural model, trained on link prediction, with competitive scores. Moreover, we used these features to train a node classifier that would distinguish troll accounts from real ones. The results are part of an ongoing project and will be finalized soon.

In the future, we consider important to leverage other types of intrinsic information that comes inherent within social media. For instance, using the actual tweeted text might give good insights to improve our presented accuracies. Even more challenging, we consider necessary to acquire knowledge from visual cues, such as images and videos posted online, as they might be an important explanatory variable to explain viral phenomena.

References

[1] P. Barberá. Social media, echo chambers, and political polarization. 2020.

[2] Y. Dong, N. V. Chawla, and A. Swami. Metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 135–144, New York, NY, USA, 2017. Association for Computing Machinery.

[3] E. Ferrara. #covid-19 on twitter: Bots, conspiracies, and social media activism. *ArXiv*, abs/2004.09531, 2020.

[4] E. Ferrara. What types of covid-19 conspiracies are populated by twitter bots? *First Monday*, May 2020.

[5] X. Liang, Z. Yang, B. Wang, S. Hu, Z. Yang, D. Yuan, N. Z. Gong, Q. Li, and F. He. Unveiling fake accounts at the time of registration: An unsupervised approach. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery &; Data Mining*, KDD '21, page 3240–3250, New York, NY, USA, 2021. Association for Computing Machinery.

[6] L. Ng and K. Carley. Flipping stance: Social influence on bot's and non bot's covid vaccine stance. *ArXiv*, abs/2106.11076, 2021.

[7] M. Sayyadiharikandeh, O. Varol, K. Yang, A. Flammini, and F. Menczer. Detection of novel social bots by ensembles of specialized classifiers. *CoRR*, abs/2006.06867, 2020.

[8] K. Sharma, Y. Zhang, E. Ferrara, and Y. Liu. Identifying coordinated accounts on social media through hidden influence and group behaviours. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '21, page 1441–1451, New York, NY, USA, 2021. Association for Computing Machinery.

[9] S. Zannettou, B. Bradlyn, E. D. Cristofaro, G. Stringhini, and J. Blackburn. Characterizing the use of images by state-sponsored troll accounts on twitter. *ArXiv*, abs/1901.05997, 2019.

[10] S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 218–226, New York, NY, USA, 2019. Association for Computing Machinery.

[11] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn. Who let the trolls out? towards understanding state-sponsored trolls. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 353–362, New York, NY, USA, 2019. Association for Computing Machinery.

[12] M. Zhang and Y. Chen. Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*, pages 5165–5175, 2018.