Curating the Twitter Election Integrity Datasets for Better Online Troll Characterization

Albert M. Orozco Camacho alorozco53@mila.quebec
Reihaneh Rabbany reihaneh.rabbany@mila.quebec

School of Computer Science, McGill University
Mila - Québec AI Institute

Abstract

- In modern days, social media platforms provide accessible channels for 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** efforts.
- Secondly, we analyze how troll activity fluctuates over time, and how it compares to a control group of *real and active users*.
- We present baselines for such tasks and highlight the differences there may exist within the literature (e.g. [2]).
- Finally, we utilize the representations learned for behaviour prediction to classify trolls from "*real*" users, using a sample of non-suspended active accounts.

Dataset Generalities

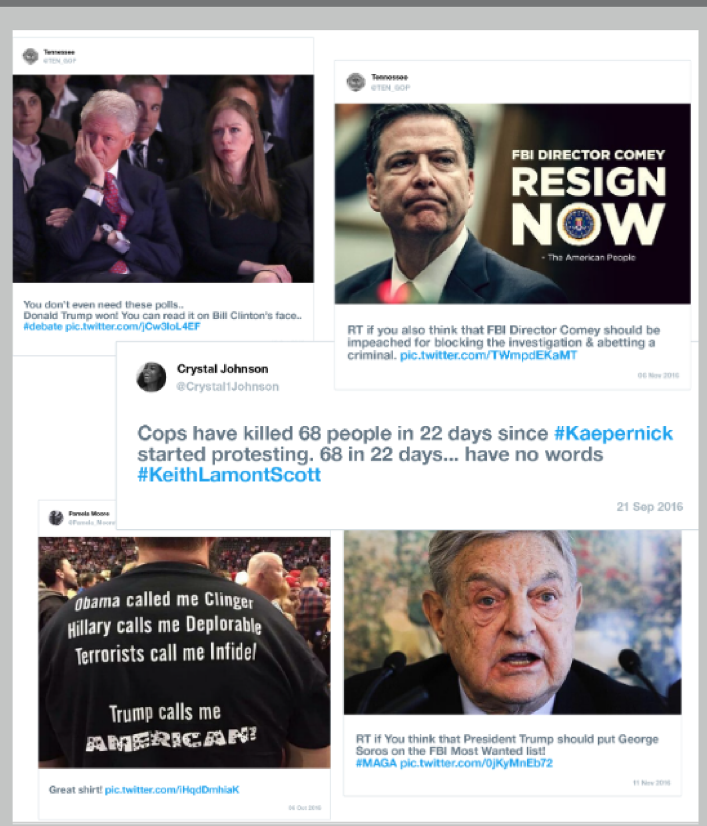

Figure 1: Sample tweets posted by accounts from the Twitter Election Integrity data set. Taken from `https://about.twitter.com/en_us/values/elections-integrity.html#data`

- The Twitter *Information Operations* database has been consistently renewed since their initial 2018 release of 4,383 accounts (see Figure 1 for sampled content).
- All released users have already being suspended. For further information, refer to `https://transparency.twitter.com/en/reports/information-operations.html`.
- Table 1 summarizes the data set statistics. We also make use of a set of **REAL** users that we have crawled during the covid-19 pandemic, for comparison purposes.

Dataset Statistics

	#senders	#receivers	#hashtags	#user mentions
Russian	129,877	1,428,207	455,853	972,354
Russian-1-hop	43,630	895,790	228,754	667,036
IRA	119,719	4,431,274	2	4,431,272
IRA-1-hop	46,551	1,237,105	396,117	840,988
Chinese	38,698	448,298	211,013	237,285
Chinese-1-hop	43,230	1,924,381	587,044	1,337,337

Table 1: Total number of *nodes* (senders, receivers), *links* (hashtags, user mentions), and *activity* (tweets) of the TEI dataset.

Network Modeling

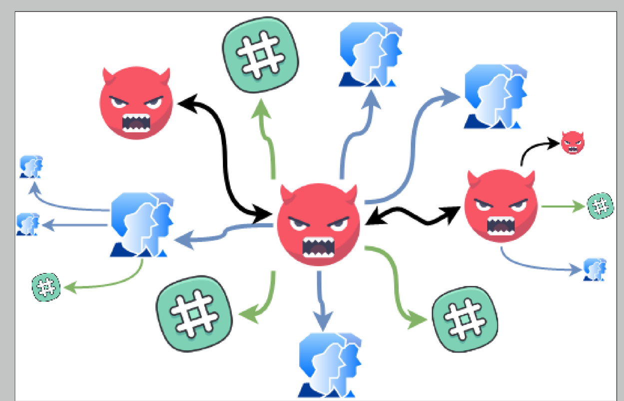

Figure 2: Trolls (in red), hashtags (in green), and real users (in blue) connect within each other via different types of links. Our modeling choice defines a heterogeneous directed graph, where trolls and real users are surrounded around their mention and hashtag relations.

- We collect a set of mentioned users by the reported trolls (1-hop neighborhood) from which we crawl their respective mention and hashtag activities.
- We extract all available tweets from a defined time frame to build a heterogeneous graph that follows the pattern depicted on Figure 2.

Methodology

- We utilize the `metapath2vec` algorithm [1] which biases random walks according to predefined node paths.
- We then use the **SEAL** [3] 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 and a `multi-layer perceptron` to do the final classification

Processing Pipeline

Algorithm 4: Node Classification Overview

Input: $G_{\mathcal{D}} = (V_{\mathcal{D}}, E_{\mathcal{D}}), A_{\mathcal{D}}, \mathbb{P}, k, Y_{\mathcal{D}}$
Output: $\bar{Y}_{\mathcal{D}}$
1 $\tilde{A}_{\mathcal{D}}$, `mask` $\leftarrow$ `negative_sampling(`$A_{\mathcal{D}}$`);`
2 $F_{\mathcal{D}} \leftarrow$ `metapath2vec(`$\tilde{A}_{\mathcal{D}}$`,` $\mathbb{P}$`);`
3 $X_{\mathcal{D}} \leftarrow$ `links2subgraphs(`$\tilde{A}_{\mathcal{D}}$`,` k`);`
4 $\bar{Z}_{\mathcal{D}} \leftarrow$ `DGCNN(`$\tilde{X}_{\mathcal{D}}$`, mask);`
5 $H \leftarrow$ `empty_tensor(dim=(`$|V_{\mathcal{D}}|$`,));`
6 **for** $v \in V_{\mathcal{D}}$ **do**
7   **for** $u \in N(v)$ **do**
8         $H[v] \leftarrow$ `concat([`$H[v]$`,` $\bar{Z}_{\mathcal{D}}[(v,u)]$`]);`
9   **end**
10     $H[v] \leftarrow$ `mean_pool(`$H[v]$`);`
11 **end**
12 $\bar{Y}_{\mathcal{D}} \leftarrow$ `mlp(`H`,` Y`, out_dim=2);`

Figure 3: This pseudo-code summarizes our deep pipeline, which can be divided in three major components: node feature computation (Lines 1-2), link prediction (Lines 3-4), and node classification (Lines 5-12).

Experiments

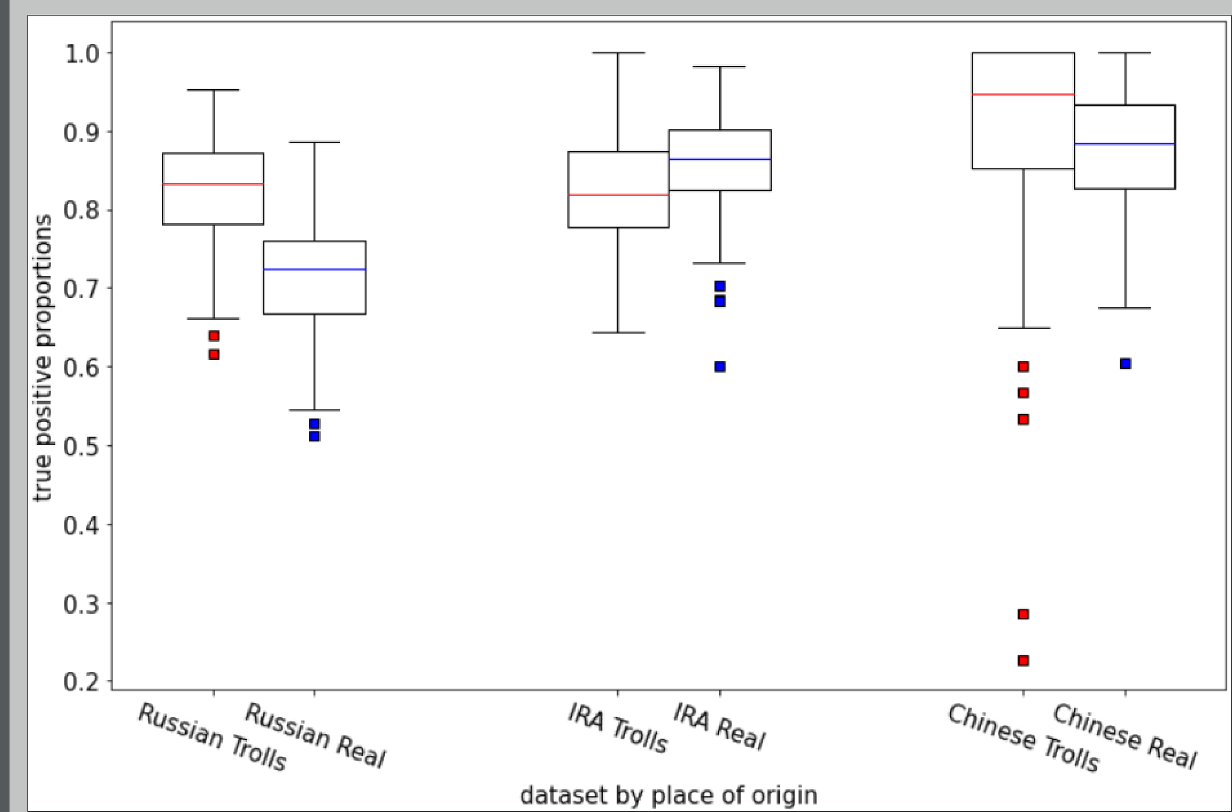

Figure 4: Proportion of correctly predicted links per each place of origin, further divided by whether each activity was produced by a troll or by a real user.

	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. We report F1-scores and accuracies averaged over every repeated experiment, defined by a sliding window over time.

Discussion and Conclusions

- To summarize, 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 *trolls* perform online.
- We were able to learn a state-of-the-art deep neural model, trained on link prediction, with competitive scores (Figure 3).
- Moreover, we used these features to train a node classifier that would distinguish troll accounts from real ones (Table 2).
- We found out that for certain group of trolls, namely those with *Russian* and *Chinese* origin, their activities (link existence distribution) is **more predictable** than those produced by a contrasting set of real accounts (Figure 4).

Future Work

- 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] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami.
metapath2vec: Scalable representation learning for heterogeneous networks.
In KDD '17, pages 135–144. ACM, 2017.

[2] Karishma Sharma, Yizhou Zhang, Emilio Ferrara, and Yan 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.

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