Discouraging Abusive Behavior in Privacy-Preserving Online Social Networking Applications

Álvaro García-Recuero [*]
Inria Rennes - Bretagne Atlantique
Rennes, France
alvaro.garcia-recuero@inria.fr

ABSTRACT

In this position paper we present the challenge of detecting abuse in a modern Online Social Network (OSN) while balancing data utility and privacy, with the goal of limiting the amount of user sensitive information processed during data collection, extraction and analysis. While we are working with public domain data available in a contemporary OSN, our goal is to design a thorough method for future alternative OSN designs that both protect user's sensitive information and discourage abuse.

In this summary, we present initial results for detecting abusive behavior on Twitter. We plan to further investigate the impact of reducing input metadata on the quality of the abuse detection. In addition, we will consider defeating Byzantine behavior by opponents in the system.

Keywords

Online Social Networks; Abuse Detection; Privacy preservation; Machine Learning

1. INTRODUCTION

The main challenge of our working proposal is applying machine learning heuristics to detect and ultimately discourage abusive behavior in OSNs, maximizing utility from a minimum amount of data.

Existing OSNs suffer from abusive behavior by their participants, who are able to use OSNs to deny, disrupt, degrade and deceive others, in occasions having a non-negligible impact on retail services, governments credibility or even stock markets [3]. Consequently, Twitter has recently taken action by introducing changes on its user policy in an attempt to resolve the pressing issue of abuse.

Several studies have defined cyber-bullying as the act of harassing another person via any form of digital communications. This behavior is intended to harm the self-esteem or image of the target victim [22] [12]. An Internet *"troll"* or cyber-troll is someone who according to [18], is member of an online community, posts abusive comments at worst or divisive information at best to create controversy.

To address such abuse, some OSN providers employ staff to search and analyze abuse related incidents. In the Twitter environment, deployed manual solutions include tools like the TwitterBlockChain[1], a browser plugin that filters abusive messages by blocking followers directly from the OSN web interface. This is an attempt to enforce OSN usage guidelines through crowd-sourcing. However, regardless of whether abuse detection is done by staff or the crowd, any kind of manual search and filtering is time consuming and thus costly.

Therefore, intelligent systems such as the Facebook immune system (FIS) [17] are built to support the task of automating abuse detection. The FIS system relies on information from user activity logs to automatically detect and act upon suspicious behavior in the OSN.

According to The Verge, Twitter is planning to collect cell phone numbers of reported users[2], presumably to create a ground-truth database of Internet trolls, where Twitter would be able to confirm recurrent abuse from these users, thus permanently suspending their accounts. However, it is yet to be proven if this alone will discourage abusive behavior completely, as abusive users can create fresh accounts in order to start abusing again.

Such automated or semi-automated methods are not perfect. For example, for the FIS, [4] found that only about 20% of the deceitful profiles they deployed were actually detected, showing that such methods result in a significant number of false negatives.

The Sybil attack [10], whereby a single user creates many identities, is a well-known theoretical tool for deceptive attacks. The research community has discussed various social-graph-based Sybil defenses [24, 20, 21, 26, 25]. However, there is little evidence of wide industrial adoption of these techniques.

While improving abuse detection remains a challenge by itself, our work is additionally concerned with preserving users' privacy. Existing OSNs collect sensitive private data from their participants for advertising[3], thus violating reasonable privacy expectations of citizens, and to the point

[*] supervised by Christian Grothoff

WWW 2016 Companion, April 11–15, 2016, Montr al, Qu bec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://d .doi.org/10.1145/2872518.2888600.

[1] https://chrome.google.com/webstore/detail/twitter-block-chain/dkkfampndkdnjffkleokegfnibnnjfah?hl=en

[2] http://www.theverge.com/2015/2/26/8116645/twitter-improves-abuse-reporting-tools-phone-numbers

[3] http://www.theguardian.com/technology/2015/mar/18/twitter-puts-trillions-tweets-for-sale-data-miners

that these private companies have become critical keystone in the military-industrial espionage complex [16].

We expect that privacy will play a bigger role in the design of future decentralized OSNs, and thus abuse detection methods should also be designed to respect privacy requirements. Decentralized OSNs would not even have an administrator to continuously verify user reports and suspend abusive accounts. Thus, in the future abuse should be automatically detected and integrated into the process that decides on whether to display an event to the end user. To evaluate how such methods may work in future OSNs, we evaluate possible methods on Twitter today. Ironically, this is enabled by the fact that on Twitter the users do not enjoy the privacy that we envision for future OSNs.

2. STATE OF THE ART

2.1 Detection of dishonest behavior

Previous work on abuse detection employs multimedia analysis techniques with the support of text-, audio-, and video-analysis to detect inappropriate content or behavior [23]. While possible to use priorly defined rules for such task, nowadays most of these algorithms learn these rules from large *corporas* of real-world examples [4].

Such techniques are complex and do not benefit from the structured communication patterns found in OSNs. Furthermore, they largely rely on features extracted from cleartext in the communications, which limits their use in applications that need to provide confidential communication.

To deal with abuse in an OSN without considering the content of the communication, graph-based research methods are emerging as an alternative to traditional text based approaches using Natural Language Processing (NLP). Graph-based techniques have been shown to be useful for detecting and combating dishonest behavior [15] and cyber-bullying [11], as well as to detect fake accounts in OSNs [7].

Graph-based methods may benefit from machine learning techniques using social-graph metadata in their feature set, for example to detect fake accounts as in [5]. Also, based on gender classification of Twitter profiles, [1] investigated the detection of deceptive profiles by looking at profile attributes such as first names, profile-, text-, link- and sidebar-colors.

The classification problems that arise in this context are often characterized by uncertainty, as for example it may never be clear whether a message is really abusive or an account is really fake. Belief function theory is used to solve problems with uncertain, incomplete or even missing data. For example, in [9], authors apply belief function theory to the problem of *"troll"* detection.

2.2 Discouraging abusive behavior

Once abusive behavior has been identified, the next logical step is to take action based on the classification. Censorship or criminal prosecution are possibilities, but such methods may be inappropriate for automatic classifiers that may have a significant number of false-positives.

The authors of [14] introduce a platform which relies on a credit-based messaging framework to make sure links among users reflect the nature of their communications. The idea is to make users more cautious about sending non-acceptable content or *spammy* requests to other users by associating a cost to establishing links.

3. PROPOSED APPROACH

To describe our approach, we begin by trying to give a definition of what constitutes abusive behavior. We then present a formal model of the OSN and the data we collect from the OSN. We then explain how we manually annotated the collected data to obtain a ground-truth database useful for training and evaluating learning algorithms.

3.1 Abuse theory

To establish whether an action is abusive, we propose a simple set of guidelines that seems to cover the various guidelines and descriptions of abusive behaviors that we found in the literature. Our definition of what constitutes abuse is based on the description of professional infiltration guidelines used by British government's Joint Threat Research and Intelligence Group's (JTRIG) [13]. JTRIG's goal is to assist the UK government in manipulating foreign populations using a combination of behavioral sciences and online warfare.

JTRIG's staff characterizes their HUMINT operations using what they call the four *D's*, and we use the same four *D's* as our definition of what kind of behavior should be considered abusive:

- Deny: encouraging self-harm to others users, promoting violence (direct or indirect), terrorism or similar activities.
- Disrupt: distracting provocations, denial of service, flooding with messages, promote abuse.
- Degrade: disclosing personal and private data of others without their approval as to harm their public image/reputation.
- Deceive: spreading false information, including supplanting a known user identity (impersonation) for influencing other users behavior and activities, or assuming false identities (but not pseudonyms that are recognizable as such).

3.2 OSN model

We will consider two directed graphs whose set of vertices are Twitter users (Figure 1). Let $\mathcal{G}_f$=$(\mathcal{V}_f, \mathcal{E}_f)$ be a directed graph of social relationships, with a set of vertices $\mathcal{V}_f$ which are follower users (those who follow or subscribe to other user's posts), and a set of directed edges $\mathcal{E}_f$ pointing from follower to followee users (those who receive such a follow or subscription request).

Secondly, let $\mathcal{G}_m$=$(\mathcal{V}_m, \mathcal{E}_m)$ be a directed messaging graph with a set of users as vertices $\mathcal{V}_m$, and a set of directed edges $\mathcal{E}_m$. These edges are created from tweets in two cases: First, they point from users authoring a tweet to users mentioned in the tweet (`@user`). Second, if a tweet is a reply, an edge is created so that it points from the responding user to the author of the original tweet. Thus, $\mathcal{E}_m$ models the tweets that are shown in a user's notifications and are thus a vector for abusive behavior. Users in set $\mathcal{V}_m$ may or may not be in the set of follower users $\mathcal{V}_f$.

For each user $u \in (\mathcal{V}_f \cap \mathcal{V}_m)$, we note the direction of its graph relationships. If it is a follow belonging to the edge set

[4] https://en.wikipedia.org/wiki/Corpus_linguistics

$\mathcal{E}_f$ of the social graph $\mathcal{G}_f$, then $\mathcal{E}_f \coloneqq \{(u,v) \mid u \text{ follows } v\}$. And if a tweet within the set of tweets $\mathcal{E}_m$ in the messaging graph $\mathcal{G}_m$, then $\mathcal{E}_m \coloneqq \{(u,v) \mid u \text{ mentions } v \vee u \text{ has reply to } v\}$.

For each tweet $\chi_i \in \mathcal{E}_m$ in the messaging graph $\mathcal{G}_m$, our classifier will define a feature (f) based binary oracle function O_f, as to predict whether the tweet in question (χ_i) is abusive or not. That is, whether it belongs to a set of acceptable tweets A, or another set of abusive tweets B. The classifier is not allowed to output "undecided", hence A∩B $= \emptyset \wedge$ A∪B $\in \mathcal{G}_m$.

3.3 Data Collection

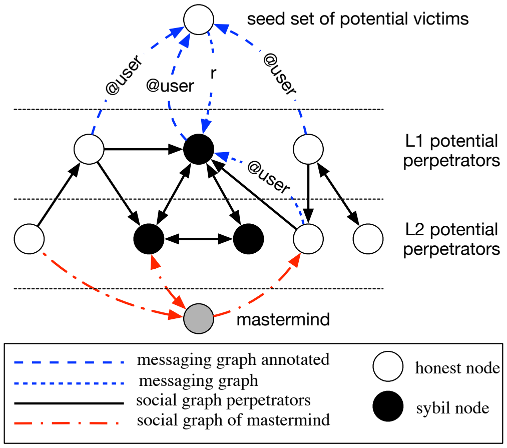
In order to be able to perform abuse detection, we have collected a database of tweets from Twitter using their public API service. We manually annotate a subset of those to create a ground-truth database.

To ensure that a relevant amount of abuse is present in the annotated sample of tweets, we started by manually identifying a few user profiles, namely *potential victims.* We start with a few profiles which are *very likely victims* of abuse on Twitter, based on the fact that they were politically active. Additionally, we selected a few accounts at random. Together these accounts are the *seed set of potential victim* accounts used during data collection. We also ensured that all of these seed accounts were largely active in English, so that we could hope to comprehend the interaction. Finally, we made sure that the accounts did not receive an excessive number of messages, as "celebrities" may be easy to manually identify as likely victims, but would likely not be representative of the whole population, and would have also caused excessive manual annotation work.

We then collected a *messaging graph* which consists of all public messages directed towards accounts in the seed set. While abuse may also occur in private messages, our API-based access to Twitter would not allow us to observe such messages. Thus, the messaging graph contains public tweets *mentioning* accounts (`@user`) in the seed set, as well as public *replies* to tweets written by the seeds. The authors of messages in the messaging graph that mention accounts in the seed set are *potential first-degree perpetrators*, in the same sense that the seeds are *potential victims.*

To allow our algorithm to detect communication patterns among the potential perpetrators, we also collected the social graph (friends, followers) and all public messages exchanged by the potential first degree perpetrators. The assumption here is that this information would be useful if multiple perpetrators use the same social network to coordinate their activities, or if a perpetrator has setup sock puppet accounts (or Sybils) and left structural evidence of their artificial creation in the data. Accounts interacting with potential first degree perpetrators (other than the seed) are considered *potential second degree perpetrators.* To avoid capturing entirely unrelated data, we did not generally collect information about verified accounts, accounts with high number of follower users or potential third degree perpetrators, as this would result in an excessively large part of the social graph being collected, which is also unlikely to be related to abusive activity. While we have no data to support the theory that popular people do not typically show abusive behavior under their real name, we do not see how excluding such atypical accounts would introduce an undue bias. Excluding these atypical accounts suppressed collection of only 40 accounts out of roughly 400,000 profiles collected belonging to the social graphs of the remaining included users. While we generally did not look beyond second degree for those graphs, if the same user interacted with a "minimum", say 3, potential second degree perpetrators, we include the account in the crawl as a *potential mastermind.* If masterminds turn out to be a useless concept, we will restrict our analysis to consider only second degree perpetrators in the future.

Figure 1: Social and messaging graphs between a mastermind and a potential victim


seed set of potential victims
@user
@user
r
@user
@user
L1 potential perpetrators
L2 potential perpetrators
mastermind
messaging graph annotated
messaging graph
social graph perpetrators
social graph of mastermind
honest node
sybil node

3.4 G round-truth database

We have created Trollslayer [5], a tool that allows volunteers to manually annotate collected tweets from Twitter. We enlisted various friends and colleagues to assist with the annotation effort, thereby hopefully obtaining volunteers capable of performing the task according to the guidelines.

We only show to volunteers the tweets in $(\mathcal{E}_m)$ that are directed to the seeds and ask them to follow the set of guidelines from Section 3.1. In addition to the tweet itself, we display some supporting context, such as previous and subsequent tweets in the tweet's author timeline. The goal is to help volunteers understand the context of a particular tweet. However, we obscure the account names involved in the conversation to minimize exposing private information and to elicit annotations that are specific to the tweet and not the author.

Volunteers provide the tool a non-binary classification of tweets. They can label a tweet as *acceptable*, *abusive* or *undecided.* The latter option is important as even with relatively clear guidelines, humans are often unsure if a particular tweet is abusive, especially given a limited context. To further offset this uncertainty, each tweet is annotated multiple times by independent volunteers.

We take a conservative approach when aggregating results of annotations, and rely on a quorum or simple majority vote from users annotating a given tweet in order to consider it really abusive. Otherwise, we cannot conclude it really represents abuse. In the latter case, it will be marked either as acceptable or undecided, according to a defined threshold value. We tune such threshold value in order to make sense

[5] http://trollslayer.decentralise.rennes.inria.fr

of annotations. To the best of our knowledge, ours is the first research that aims to create a ground-truth database that associates tweets with a clear definition of abuse.

4. METHODOLOGY

The goal of our algorithm is to rate the abusiveness of tweets from the messaging graph. The resulting ranking would then be used by a personalized filter operating on behalf of the potential victim.

Given a ground-truth database with tweets annotated as *acceptable*, *abusive* or undecided, we plan to train and evaluate learning algorithms on aggregated information from the given annotations. This is similar to what work for credibility analysis on Twitter has previously shown to work well for building models on topic classification [8].

The collected dataset will be partitioned into a set for training of the machine learning/classification algorithm(s), and a disjoint set for the evaluation. The partitions are induced by the seeds during our initial data collection as explained in Section 3.3. In particular, all collected information relating to the same seed remains in the same partition.

We first plan to evaluate a few off-the-shelf supervised models [6], linear ones as for instance Bayesian Regression. Then, Naive Bayes algorithms as Multinomial Naive Bayes, or Decision Tree algorithms: ID3, C4.5, C5.0 and CART [2]; all while fitting our proposed methodology and set of features. Later, we might attempt to combine several classifiers and check if that yields a better solution to the problem of abuse detection.

Our oracle function will use these algorithms and the set of features extracted in section 4.1 in order to output a decision for each tweet collected.

4.1 Feature engineering

In order to build an abuse classifier, we apply a number of transformations on the data and extract a set of features to input into learning models. Then, we can evaluate how different models perform.

Within our OSN model, features fall into four categories: tweet, user, social graph and messaging graph related data.

For tweets, we for example extract numerical features such as the number of mentions, hashtags in a tweet directed to the seed set. In the case of user accounts, we present a couple of them here, such as follower and followee count. For the messaging graph, we outline mentions and their replies. As for social graph-related data, the mutual followers and/or mutual followees among the seed and the authors of the tweets with mentions we collect for annotation may be eventually important to analyze the connectivity of a given tweet author in the social graph, and therefore added too.

Within our model, we plan to record and assess the relative importance (RI) of each of the features within the categories described. For instance, the random forest (RF) learning algorithm [6] will output a value to highlight the relative importance of each feature during the decision making process (classify as abusive or not).

4.2 Classifier metrics

The output of the abuse classifier will depend on a given *threshold*, which is a cutoff value in the prediction probability after which the classifier identifies a tweet as potentially abusive. In order to capture the trade-off between true positive rate (TPR) and false positive rate (FPR) in a single curve, the receiver operating characteristics (ROC) analysis offers the possibility of visualizing the trade-off resulting from different *threshold* values.

In ROC, the closer the curve is to the upper left corner at coordinate (0, 1) the better the classifier performance is. Therefore, the quality of the classifier predictions can be assessed by calculating the area under the curve (AUC). However, given that abusive messages are expected to be rare, precision-recall curves might be a more appropiate method than ROC curves.

5. EVALUATION

Table 1 shows some basic statistics about the collected data set, such as the number of seeds and tweets collected for annotation. We provide the number of tweets annotated as abusive, counting only those obtained with the majority voting scheme mentioned. For instance, if two reviewers out of three mark a tweet as abuse, then it is abuse. Same if we have 7 reviewers, then we would need 4 to agree a tweet is really abusive.

In total, there are 1648 tweets with mentions, out of which 30 have been confirmed to be abusive, and 539 acceptable, based on a majority vote calculated among three different volunteers.

Table 1: Table reflecting stats of abuse annotated

	Total	Abuse
Seed set of potential victims	47	
# very likely victims	10	
# random accounts	37	
$\mathcal{E}_m \in \mathcal{G}_m$ from L1 to seed set	1648	
with mention/s	1648	30
with mention/s & replies	428	13
$\mathcal{V}_m \in \mathcal{G}_m$ L1 potential perpe.	1113	16

Table 2 shows the value distribution for various features related to tweets and their authors. It lists the cummulative distribution functions (CDFs) of a few features only, due to space limits, but we have more. In the X axis we represent the value range normalized in respect to the maximum value of the feature with all annotated tweets, while the Y axis represents the fraction of tweets above the respective thresholds (all, acceptable, abusive).

6. FUTURE WORK

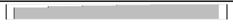
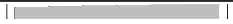
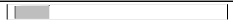
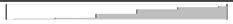
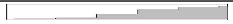
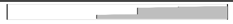
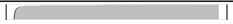
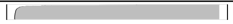
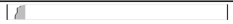
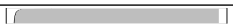
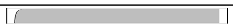
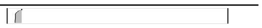
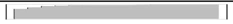
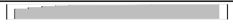
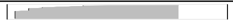
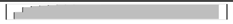
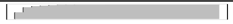
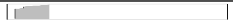
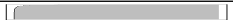
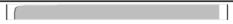
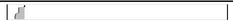
The next step is to evaluate various machine learning methods to detect abusive behavior based on the features we have extracted. Given the clear distinctions in the CDF patterns, we are optimistic that learning methods will produce good results.

A user's social graph is still sensitive personal information, hence we plan to investigate the use of secure multiparty computations and other data minimization techniques to reduce the amount of sensitive information that needs to be exposed in order to successfully detect abusive behavior.

Another possible direction would be to generalize our techniques to other OSN models; however, our current data set

[6] http://scikit-learn.org/stable/supervised_learning.html

Table 2: CDFs of a sample list of features for annotated tweets

	Overall	Acceptable	Abusive
#hashtags in tweet to a potential vic.			
#mentions in tweet to a potential vic.			
#followers in user account			
#followees in user account			
#reply tweets to/from potential vic.			
ratio #(followees/followers)			
ratio #(followers/followees)			

is limited to Twitter and it would be rather costly to also obtain message data and ground truth for other OSNs.

Eventually, we also plan to integrate our results with an emerging decentralized OSN [19] to create a privacy-preserving OSN where abusive behavior is discouraged. However, instead of establishing a credit system, we plan to simply rank items on the receiving side, assigning those that are abusive, irrelevant or undesired a lower chance of being displayed.

7. REFERENCES

[1] J. S. Alowibdi, U. Buy, P. S. Yu, L. Stenneth, et al. Detecting deception in online social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 383–390. IEEE, 2014.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*, volume 4. 2006.

[3] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011. http://dx.doi.org/10.1016/j.jocs.2010.12.007.

[4] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: When bots socialize for fame and money. In *Proceedings of the 27th Annual Computer Security Applications Conference*, ACSAC '11, pages 93–102, New York, NY, USA, 2011. ACM.

[5] Y. Boshmaf, M. Ripeanu, K. Beznosov, and E. Santos-Neto. Thwarting fake osn accounts by predicting their victims. In *Proceedings of the 8th ACM Workshop on*