Mean Birds: Detecting Aggression and Bullying on Twitter

Despoina Chatzakou[†], Nicolas Kourtellis[‡], Jeremy Blackburn[‡], Emiliano De Cristofaro[♯], Gianluca Stringhini[♯], and Athena Vakali[†]

[†]Aristotle University of Thessaloniki [‡]Telefonica Research [♯]University College London

deppych@csd.auth.gr, nicolas.kourtellis@telefonica.com, jeremy.blackburn@telefonica.com
e.decristofaro@ucl.ac.uk, g.stringhini@ucl.ac.uk, avakali@csd.auth.gr

Abstract

In recent years, bullying and aggression against users on social media have grown significantly, causing serious consequences to victims of all demographics. In particular, cyberbullying affects more than half of young social media users worldwide, and has also led to teenage suicides, prompted by prolonged and/or coordinated digital harassment. Nonetheless, tools and technologies for understanding and mitigating it are scarce and mostly ineffective.

In this paper, we present a principled and scalable approach to detect bullying and aggressive behavior on Twitter. We propose a robust methodology for extracting text, user, and network-based attributes, studying the properties of cyberbullies and aggressors, and what features distinguish them from regular users. We find that bully users post less, participate in fewer online communities, and are less popular than normal users, while aggressors are quite popular and tend to include more negativity in their posts. We evaluate our methodology using a corpus of 1.6M tweets posted over 3 months, and show that machine learning classification algorithms can accurately detect users exhibiting bullying and aggressive behavior, achieving over 90% AUC.

1 Introduction

Cyberbullying and cyberaggression are increasingly serious and widespread issues affecting large numbers of Internet users. In the physical world, bullying entails repeated negative/aggressive interactions, e.g., in the form of threats, damaging rumors, and verbal or physical attacks. Its digital manifestation, cyberbullying, differs in several ways, as harassment may be carried out by total strangers, also potentially "protected" by anonymity. Moreover, today's hyperconnected society allows a phenomenon once limited to particular places or times of the day (e.g., school hours) to occur anytime, anywhere, with just a few taps on a keyboard. Cyberbullying and cyberaggression can be defined in different ways [19, 46, 51] – we denote *cyberbullying* as the repeated and hostile behavior by a group or an individual, using electronic forms of contact, and *cyberaggression* as intentional harm delivered via electronic means to a person or a group of people who perceive such acts as offensive, derogatory, harmful, or unwanted [20].

Over the past few years, as social interactions have increasingly moved online, social media has become an integral part of people's lives. At the same time, however, bullying and aggression against Internet users has also increased. Only a few years ago, cyberbullying was not taken seriously: the typical advice was to "just turn off the screen" or "disconnect" [35]. However, as its proliferation and the extent of its consequences reach epidemic levels [3], this behavior can no longer be ignored: in 2014, about 50% of young social media users reported being bullied online in various forms [16]. Popular social media platforms like Twitter and Facebook are not immune [44], as racist and sexist attacks may even have caused potential buyers of Twitter to balk [50].

Despite the seriousness of the problem, there are very few successful efforts to detect abusive behavior on Twitter, both from the research community (see Section 6) and Twitter itself [49]. Arguably, there are several inherent obstacles. First, tweets are short and full of grammar and syntactic flaws, which makes it harder to use natural language processing tools to extract text-based attributes and characterize user interactions. Second, each tweet provides fairly limited context, therefore, taken on its own, an aggressive tweet may be disregarded as normal text, whereas, read along with other tweets, either from the same user or in the context of aggressive behavior from multiple users, the same tweet could be characterized as bullying. Third, despite extensive work on spam detection in social media [18, 47, 56], Twitter is still full of spam accounts [8], often using vulgar language and exhibiting behavior (repeated posts with similar content, mentions, or hashtags) that could also be considered as aggressive or bullying. Filtering out such accounts from actual abusive users may be a difficult task. Finally, aggression and bullying against an individual can be performed in several ways beyond just obviously abusive language – e.g., via constant sarcasm, trolling, etc.

Overview & Contributions. In this paper, we tackle the problem of detecting cyberbullying and aggressive behavior on Twitter. We design and execute a novel methodology geared to label aggressive and bullying behavior of Twitter users. Specifically, by presenting a principled and scalable approach for eliciting user, text, and network-based at-

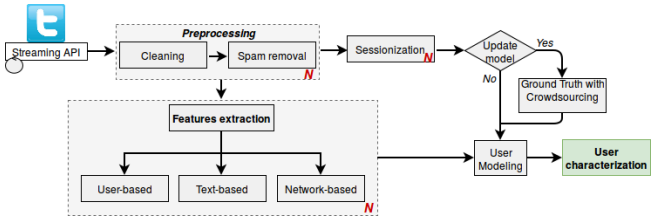


Figure 1: Overview of our methodology. N denotes the ability to parallelize a task on N processors.

tributes of Twitter users, we extract a total of 30 features. We study the properties of cyberbullies and aggressors, and what features distinguish them from regular users, alongside labels provided by human annotators recruited from CrowdFlower [11], a popular crowdsourcing platform. Such labels, contributed infrequently or at regular intervals, can be used to enhance an already trained model, bootstrapping the detection method and executed on large set of tweets.

We experiment with a corpus of $1.6M$ tweets, collected over 3 months, finding that bully users are less "popular" and participate in fewer communities. However, when they do become active, they post more frequently, and use more hashtags, URLs, etc., than others. Moreover, we show that bully and aggressive users tend to attack, in short bursts, particular users or groups they target. We also find that, although largely ignored in previous work, network-based attributes are actually the most effective features for detecting aggressive user behavior. We show that our features can be fed to classification algorithms, such as Random Forests, to effectively detect bullying and aggressive users, achieving up to 0.907 weighted AUC (Area Under the Receiver Operating Characteristic curve) [22], 89.9% precision and 91.7% recall. Finally, we discuss the effectiveness of our methods by comparing results with the suspension and deletion of accounts as observed in the wild for users who, though aggressive, remain seemingly undetected by Twitter.

Paper Organization. The rest of the paper is organized as follows. The next section provides a high-level overview of our methodology, then, in Section 3, we present our dataset and the steps taken for cleaning and preparing it for analysis. Section 4 discusses the features extracted from the dataset and the classes of users we consider, while Section 5 presents the techniques used to model the data and predict user behavior. After reviewing related work in Section 6, we conclude in Section 7.

2 Methodology

Our approach to detect aggressive and bullying behavior on Twitter, as summarized in Figure 1, involves the following steps: (1) data collection, (2) preprocessing of tweets, (3) sessionization, (4) extracting user-, text- and network-level features, (5) user modeling and characterization, and (6) classification.

Data Collection. Our first step is to collect tweets and, naturally, there are a few possible ways to do so. In this paper, we rely on Twitter's Streaming API, which provides free access to 1% of all tweets. The API returns each tweet in a JSON format, with the content of the tweet, some metadata (e.g., creation time, whether it is a reply or a retweet, etc.), as well as information about the poster (e.g., username, followers, friends, number of total posted tweets).

Preprocessing. Next, we remove stop words, URLs, and punctuation marks from the tweet text and perform normalization – i.e., we eliminate repeated letters and repetitive characters; e.g., the word "yessss" is converted to "yes". This step also involves the removal of spam content, which can be done using a few different techniques relying on tweeting behavior (e.g., many hashtags per tweet) and/or network features (e.g., spam accounts forming micro-clusters) [18, 56].

Sessionization. Since analyzing single tweets does not provide enough context to discern if a user is behaving in an aggressive or bullying way, we group tweets from the same user, based on time clusters, into *sessions*. This allows us to analyze contents of sessions rather than single tweets.

Ground Truth. We build ground truth (needed for machine learning classification, explained next) using human annotators. For this we use a crowdsourced approach, by recruiting workers who are provided with a set of tweets from a user, and are asked to classify them according to predefined labels. If such an annotated dataset is already available, this step can be omitted.

Feature Extraction. We extract features from both tweets and user profiles. More specifically, features can be user-, text-, or network-based; e.g., the number of followers, tweets, hashtags, etc. The selection of appropriate features is obviously a very important step to speed up and improve learning quality [30].

Classification. The final step is to perform classification using the (extracted) features and the ground truth. Naturally, different machine learning techniques can be used for this task, including probabilistic classifiers (e.g., Naïve Bayes), decision trees (e.g., J48), ensembles (e.g., Random Forests), or neural networks.

Scalability and Online Updates. An important challenge for our methodology to address is supporting *scalable* analysis of large tweet corpora. Obviously, several of the steps above can be performed in parallel, e.g., over N subsets of the data, on N cores (as depicted in Figure 1). Also, depending on whether data is processed in batches or in a streaming fashion, one can use different modeling algorithms and processing platforms – e.g., batch platforms like Hadoop vs. distributed stream processing engines like Storm. Either way, some of the steps, such as the annotation from crowdworkers, can be periodically executed on new data, and the model updated to handle changes in data and/or manifestation of new aggressive behaviors.

We argue that our pipeline design provides several benefits with respect to performance, accuracy, and extensibility,

since it allows regular updates of the model, thus capturing previously unseen human behaviors. Moreover, we can easily plug in new features, e.g., as new metadata is made available from the Twitter platform, or from new research insights. Finally, different components can be updated or extended with new technologies, e.g., allowing for better data cleaning, feature extraction, and modeling.

3 Dataset & Ground truth

In this section, we present the data used in our evaluation, and the way we process it to build ground truth.

3.1 Data Collection

Data collection took place between June and August 2016, aiming to gather two sets of tweets:

1. ***(Baseline)*** A baseline of 1M random tweets;
2. ***(Hate-related)*** A set of 650K tweets collected from the Twitter Streaming API using 309 hashtags associated with bullying and hateful speech.

More specifically, we build the list of 309 hashtags as follows: we obtain a 1% sample of all public tweets from June to August 2016, using the Streaming API, and parse it to select all tweets containing #GamerGate. The Gamergate controversy [32] is one of the most well documented large-scale instances of bullying/aggressive behavior that we are aware of, originating from alleged improprieties in video game journalism quickly growing into a larger campaign centered around sexism and social justice [4].

With individuals on both sides of the controversy using it, and extreme cases of cyberbullying and aggressive behavior associated with it (e.g., direct threats of rape and murder), #GamerGate serves as a relatively unambiguous hashtag associated with tweets that are likely to involve the type of behavior we are interested in. We use #GamerGate as a seed for a sort of snowball sampling of other hashtags likely associated with cyberbullying and aggressive behavior; we also include tweets that have one of the 308 hashtags that appeared in the same tweet as #GamerGate. Indeed, when manually examining these hashtags, we see that they contain a number of hateful words or hashtags, e.g., #IStandWithHateSpeech, #KillAllNiggers, and #InternationalOffendAFeministDay.

Apart from the hate-related set, we also crawl a random set of tweets which serves as a baseline, as it is less prone to contain any kind of offensive behaviors. As noted in our previous work [4], there are significant differences among the two sets. Thus, we consider the number of followers, the usage of hashtags in users' posts, and the expressed sentiment to highlight such differences. Figures 2a and 2b show that there are differences between the users' social and tweeting activity. We observe that users from the hate-related set have more followers compared to the baseline set. This could be because users with aggressive behavior tend to accumulate more popularity in their network (Figure 2a). Also, baseline users tweet with fewer hashtags than users from the hate-related dataset (Figure 2b), perhaps as the latter use Twitter as a rebroadcasting mechanism aiming at attracting attention to the topic. Finally, Figure 2c shows that the hate-related dataset contains more negative sentiment, in line with the fact that it contains a large proportion of offensive posts.

As expected, the hate-related dataset also contains activity of users who may not be aggressive or hateful, and in fact, may be driving metrics such as popularity in skewed ways. Therefore, to understand nuanced differences between aggressive or hateful users with normal users, we investigate in more detail behavioral patterns observed in this dataset in Section 4. More specifically, we analyze user behavior and profile characteristics based on labels (normal, bully, or aggressive) provided by human annotators.

3.2 Preprocessing

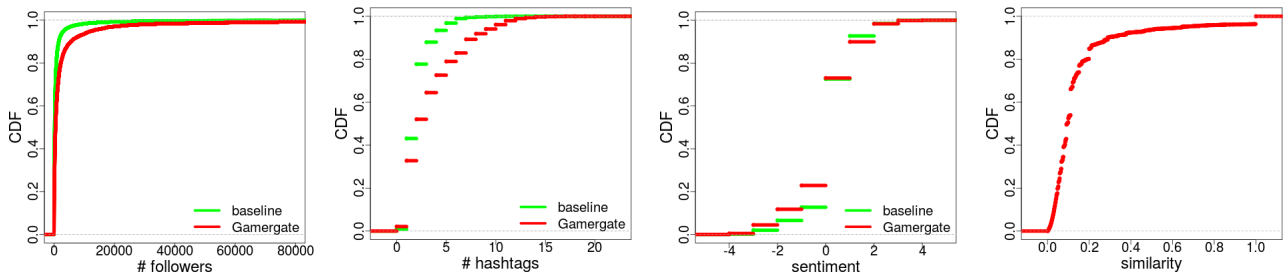
We perform several steps to prepare the data for labeling and build ground truth.

Cleaning. The first step is to clean the data of noise, i.e., removing URLs, numbers, stop words, emoticons and punctuations, as well as converting all characters to lower case.

Removing Spammers. Previous work has shown that Twitter contains a non-negligible amount of spammers [8], and proposed a number of spam detection tools [18, 56]. Here, we perform a first-level detection of spammers and remove them from our dataset. We use Wang et al.'s approach [56], relying on two main indicators of spam: (i) using of a large number of hashtags in the tweets (to boost visibility), and (ii) posting a large number of tweets that are highly similar to each other. To find optimal cutoffs for these heuristics, we study both the distribution of hashtags and the duplication of tweets. As for hashtag distribution, we notice that the average number of hashtags within a user's posts ranges from 0 to about 17. We experiment with different possible cutoffs and, after a manual inspection on a sample of posts, we set the limit to 5 hashtags. That is, users with more than 5 hashtags per tweets on average are removed. Next, we estimate the similarity of a user's tweets via the Levenshtein distance [37], i.e., the minimum number of single-character edits needed to convert one string into another, averaging it out over all pairs of the user's tweets. For each user's posts, we calculate the intra-tweet similarity: for a user with x tweets, we arrive at a set of n similarity scores, where $n = x(x-1)/2$. Then, we compute the average intra-tweet similarity per user: if it is above 0.8, we exclude them and their posting activity. Figure 2d shows that about 5% of users have a high percentage of similar posts and are thus removed from our dataset.

3.3 Sessionization

Cyberbullying behavior usually involves *repetitive* actions, thus, we aim to study users' tweets *over time*. Inspired by Hosseinmardi et al. [26] – who consider a lower bound



(a) Followers distribution. **(b)** Hashtags distribution. **(c)** Sentiment distribution. **(d)** Similarity distribution.

Figure 2: CDF of (a) Followers, and (b) Hashtags, (c) Avg. sentiment, and (d) Percentage of a user's posts with similarity above 0.8.

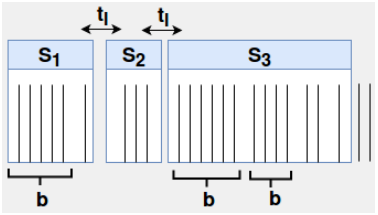


Figure 3: Sessionization and Batching process.

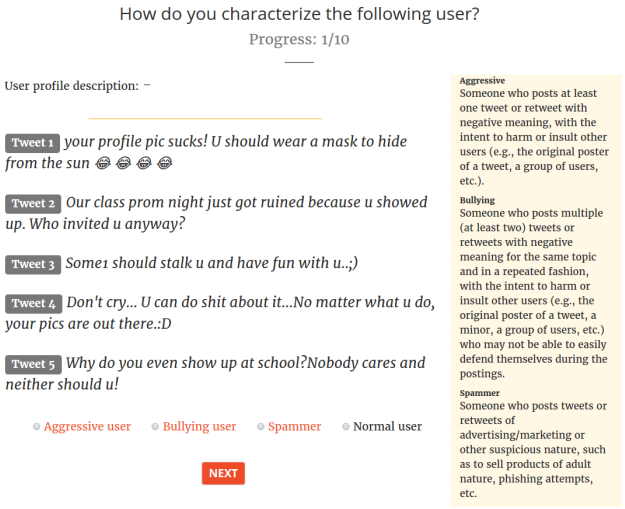


Figure 4: Example of the crowdsourcing user interface.

threshold of comments for media sessions extracted from Instagram to be presented in the annotation process – we create, for each user, sets of time-sorted tweets (sessions) by grouping tweets posted close to each other in time.

Our sessionization process is depicted in Figure 3. First, we remove users who are not significantly active, i.e., tweeting less than five times in the 3-month period. Then, we use a session-based model where, for each session S_i, the inter-arrival time between tweets does not exceed a predefined time threshold t_l. We experiment with various values of t_l to find an optimal session duration and arrive at a threshold of 8 hours. The minimum, median, and maximum length of the resulting sessions (in terms of the number of their included tweets) for the hate-related (i.e., Gamergate) dataset are, respectively, 12, 22, and 2.6K tweets. For the baseline set of tweets, they are 5, 44, and 1.6K tweets.

Next, we divide sessions in batches, as otherwise they would contain too much information to be carefully examined by a crowdworker within a reasonable period of time. To find the optimal size of a batch, i.e., the number of tweets per batch, we performed preliminary labeling runs on CrowdFlower, involving 100 workers each, using batches of exactly 5, 5-10, and 5-20 tweets. Our intuition is that increasing the batch size provides more context to the worker to assess if a poster is acting in an aggressive or bullying behavior, however, too many tweets might confuse them. The best results with respect to labeling agreement – i.e., the number of workers that provide the same label for a batch – occur with 5-10 tweets per batch. Therefore, we eliminate sessions with fewer than 5 tweets, and further split those with more than 10 tweets (preserving the chronological ordering of their posted time). In the end, we arrive at 1,500 batches. We also note that we maintain the same number of batches for both the hate-related and baseline tweets.

3.4 Crowdsourced Labeling

We now present the design of our crowdsourcing labeling process, performed on CrowdFlower.

Labeling. Our goal is to label each Twitter user – *not* single tweets – as *normal*, *aggressive*, *bullying*, or *spammer* by analyzing their batch(es) of tweets. Note that we also allow for the possibility that a user is spamming and has passed our basic spam filtering. Based on previous research [46, 51, 19], workers are provided with the following definitions of aggressive, bullying, and spam behaviors:

- *aggressive user:* someone who posts at least one tweet or retweet with negative meaning, with the intent to harm or insult other users (e.g., the original poster of a tweet, a group of users, etc.);
- *bullying user:* someone who posts multiple (at least two) tweets or retweets with negative meaning for the same topic and in a repeated fashion, with the intent to harm or insult other users (e.g., the original poster of a tweet, a minor, a group of users, etc.) who may not be able to easily defend themselves during the postings;
- *spammer user:* someone who posts texts of advertis-

ing/marketing or other suspicious nature, such as to sell products of adult nature, and phishing attempts.

CrowdFlower Task. We redirect employed crowd workers to an online survey tool we developed. First, they are asked to provide basic demographic information: gender, age, nationality, education level, and annual income. In total, 30% are female and 70% male, while their claimed educational level are spread between secondary education (18.4%), bachelor degree (35.2%), master (44%) and PhD (2.4%). One third (35.5%) claims to have an income level below €10k, and about 20% between €10k and €20k. The rest are spread in the €20k-€100k range. About 27% are 18-24 years old, 30% between 25-31 years old, 21% 32-38 years old, 12% 39-45 years old, with the remainder above 45 years old. They come from 56 different countries, with significant participation of users from USA, Venezuela, Russia, and Nigeria. Overall, the annotators from the top 10 countries contribute 75% of all annotations.

Once they have supplied demographic information, we then ask workers to label 10 batches, one of which is a control case (details below). We also provide them with the user profile description (if any) of the Twitter user they are labeling and the definition of aggressive, bullying, and spammer behaviors.

Figure 4 presents an example of the interface. The workers rated the instructions given to them, as well as the overall task, as very good with an overall score of 4 out 5.

Results. Overall, we recruit 834 workers. They are allowed to participate only once to eliminate behavioral bias across tasks and discourage rushed tasks. Each batch is labeled by 5 different workers, and, similar to previous work [26], [38], a majority vote is used to decide the final label. We receive 1,307 annotated batches, comprising 9,484 tweets in total. 4.5% of users are labeled as bullies, 3.4% as aggressive, 31.8% as spammers, and 60.3% as normal users. Overall, abusive users (i.e., bullies and aggressors) make up about 8% of our dataset, which mirrors observations from previous studies. E.g., in [28] 9% of the users in the examined dataset exhibits bad behavior, while in [1] 7% of users are cheated. Thus, we believe our ground truth dataset contains a representative sample of aggressive/abusive content.

Annotator reliability. To assess the reliability of our workers, we use (i) the Fleiss' kappa measure [17], and (ii) control cases. More specifically, Fleiss' kappa is a statistical measure which assesses the reliability of agreement between a fixed number of raters [17]. We find the inter-rater agreement to be 0.54. To evaluate the quality of the observed agreement, we estimate the $kappa$ value which shows the extent to which agreement among raters exceeds what would be expected if all raters made their ratings completely randomly. We find $kappa$ to be 0.22, which indicates a fair level of agreement between our workers [55].

Finally, we also use control cases to ensure worker "quality" by manually annotating three batches of tweets. During the annotation process, each worker is given a set of batches to annotate, one of which is a randomly selected control case: the annotation of these control cases is used to assess their ability to adequately annotate for the given task. We find 66.5% accuracy overall (i.e., the percent of correctly annotated control cases). More specifically, 83.75% accuracy for spam, 53.56% for bully, and 61.31% for aggressive control cases.

Type	Feature
User (total: 10)	avg. # posts, # days since account creation, verified account # subscribed lists, posts' interarrival time, default profile image? statistics on sessions: total number, avg., median, and STD. of their size
Textual (total: 9)	avg. # hashtags, avg. # emoticons, avg. # upper cases, # URLs avg. sentiment score, avg. emotional scores, hate score avg. word embedding score, avg. curse score
Network (total: 11)	# friends, # followers, hubs, (d=#followers/#friends), authority avg. power diff. with mentioned users, clustering coefficient, reciprocity eigenvector centrality, closeness centrality, louvain modularity

Table 1: Features considered in the study.

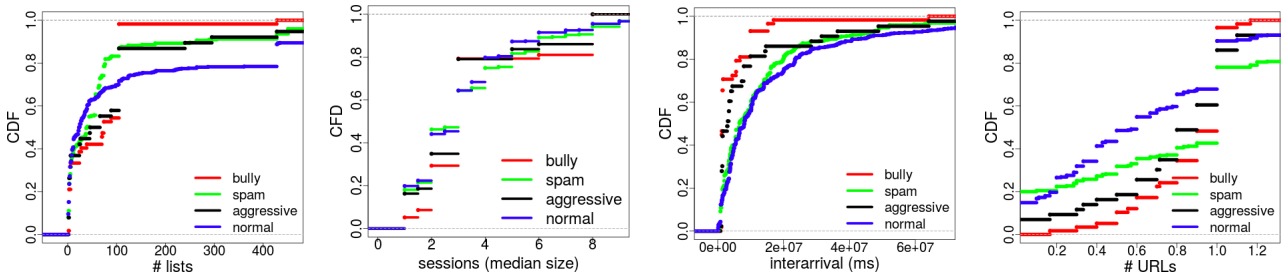
4 Feature Extraction

In this section, we focus on user-, text-, and network-based features that can be subsequently used in the modeling of user behaviors identified in the dataset. Next, we detail the features from each category. A summary is shown in Table 1. To examine the significance of differences among the distributions which are being presented next, we use the two-sample Kolmogorov-Smirnov test, a non-parametric statistical test to compare the probability distributions of different samples. We consider as statistically significant all cases with p value less than 0.05.

4.1 User-based features

Basics. We experiment with various user-based features; i.e., features extracted from a user's profile. Features in this category include the number of tweets a user has made, the age of his account (i.e. number of days since its creation), the number of lists subscribed to, if the account is verified or not (i.e., acknowledged by Twitter as an account linked to a user of "public interest"), and whether or not the user still uses the default profile image. A representative example is shown in Figure 5a, which plots the CDF of the number of subscribed lists for each of the four behaviors we examine (we note that the maximum number of lists is 4,327, but we trim the plot at 500 for readability). The median (max) number of lists for bullying, spam, aggressive, and normal users is 24 (428), 57 (3,723), 40, (1,075), and 74 (4,327), respectively. We note the difference in the participation of groups from each class of users, with normal users signing up to more lists than the other types of users.

Session statistics. Here, we consider the total number of sessions produced by a user from June to August and we estimate several statistics: average, median, and standard deviation of the size of each users' sessions. Figure 5b shows the CDF of the median number of sessions for the four user behavior classes. Comparing the distributions among the bully

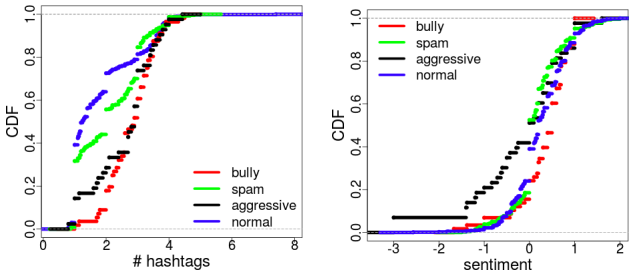


(a) Lists distribution. **(b)** Sessions distribution. **(c)** Interarrival time distribution. **(d)** URLs distribution.

Figure 5: Distribution of (a) Lists, (b) Sessions size, (c) Interarrival time, and (d) URLs.

and aggressive users to the normal users, we observe that the differences are not statistically significant with D=0.16052 and D=0.14648 for bully vs. normal, and aggressive vs. normal users, respectively.

Interarrival time. We estimate the inter-arrival time of a user's posts by considering all of his activity from June to August. From Figure 5c we observe that bullies and aggressors tend to have less waiting time in their posting activity compared to the spam and normal users, which is in alignment with the results obtained in [26].



(a) Hashtags distribution. **(b)** Sentiment distribution.

Figure 6: Distribution of (a) Hashtags, and (b) Sentiment.

4.2 Text-based features

For text-based features, we look deeper into a user's tweeting activity by analyzing specific attributes that exist in his tweets.

Basics. We consider some basic metrics across a user's tweets: the number of hashtags used, uppercase text (which can be indicative of intense emotional state or 'shouting'), number of emoticons, and URLs. For each of these, we take the average over all tweets in a users' annotated batch. Figure 5d depicts the CDF of the average number of URLs for the different classes of users. The median value for the bully and spam users is 1, for the aggressive it is 0.9, and for the normal users it is 0.6. There are differences in the maximum average number of URLs between the 4 classes: for the bully and aggressive users it is 1.17 and 2, respectively, while for spam and normal users it is 2.38 and 1.38. Thus, we see a clear separation in the tendency to post URLs. Also, from Figure 6a we observe that aggressive and bully users have a propensity to use more hashtags within their tweets, as they try to disseminate their attacking message to multiple individuals or groups.

Word embedding. Word embedding allows finding both semantic and syntactic relation of words, which permits the capturing of more refined attributes and contextual cues that are inherent in human language. E.g., people often use irony to express their aggressiveness or repulsion. Therefore, we use Word2Vec [34], an unsupervised word embedding-based approach to detect semantic and syntactic word relations. Word2Vec is a two-layer neural network that operates on a set of texts to 1) initially establish a vocabulary based on the words included in such set more times than a user-defined threshold (to eliminate noise), 2) apply a learning model to input texts to learn the words' vector representations in a D-dimensional space, and 3) output a vector representation for each word encountered in the input texts. D is user-defined, while based on [34] 50-300 dimensions can model hundreds of millions of words with high accuracy. Two methods can be used to build the actual model: 1) CBOW (i.e., Continuous bag of words), which uses context to predict a target word, and 2) Skip-gram, which uses a word to predict a target context. Skip-gram works well with small amounts of training data and handles rare words or phrases well, while CBOW shows slightly better accuracy for frequent words and is faster to train. To train Word2Vec, either large textual corpora are used (e.g., Wikipedia articles in a selected language), or more thematic textual collections to better embed the word usage in the targeted domain.

Here, we use Word2Vec to generate features to better capture the context of the data at hand. We use a pre-trained model with a large scale thematic coverage (with 300 dimensions) and apply the CBOW model due to its better performance regarding the training execution time. Finally, having at hand the vector representations of all input texts' words, the overall vector representation of an input text is derived by averaging all the vectors of all its comprising words. Comparing the bully distribution with the normal one we conclude to D=0.094269 and p=0.7231, while in the aggressive vs. normal distribution comparison D=0.11046 and p=0.7024, thus in both cases the differences are not statistically significant.

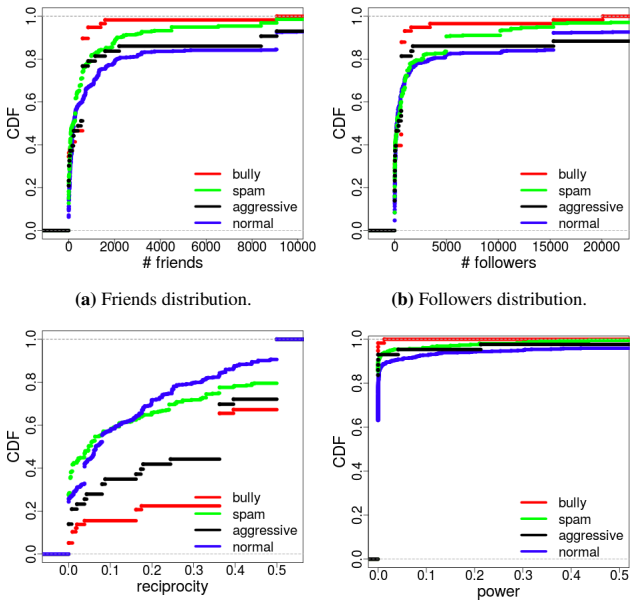
Sentiment. Sentiment has already been considered during

the process of detecting offensive or abusive behavior in communications between individuals, e.g., [36]. To detect sentiment, we use the SentiStrength tool [45], which estimates the positive and negative sentiment (on a [-4, 4] scale) in short texts, even for informal language often used on Twitter. First, however, we evaluate its performance by applying it on an already annotated dataset containing a total of $7,086$ tweets [53]. The overall accuracy is 92%, attesting to its efficacy for our purposes. Figure 6b plots the CDF of average sentiment for the 4 user classes. Even though we observe a distinction among the aggressive and the rest of classes, this is not the case when comparing the remaining three classes, where more or less similar behavioral patterns are observed concerning the expressed sentiments. More specifically, comparing the distributions of the aggressive classes with the normal, they are different with a test statistic D=0.27425, while considering the bully and normal distributions the differences are statistical significant with D=0.27425. We also attempt to detect more concrete emotions, i.e., anger, disgust, fear, joy, sadness and surprise based on the approach presented in [5]. Comparing the distributions of the abusive classes with the normal, in most of the cases we observe that there are no statistical differences. For instance, considering joy D=0.046291 when comparing the bully and normal distributions, while D=0.040277 for the aggressive and normal ones with p>0.05. For anger, even though the aggressive and normal distributions are significantly different (D=0.21515, p=0.04588) the bully and normal users are not (D=0.080029 and p=0.8795).

Hate and curse words. Additionally, we wanted to specifically examine the existence of hate speech and curse words within tweets. For this purpose, we use the Hatebase database [23], which is a crowdsourced list of hate words. Each word in the Hatebase database is additionally scored on a [0, 100] scale indicating how hateful the word is. Finally, a list of swear words [39] is also used in a binary fashion; i.e., we set a variable to true if a tweet contained any word in the list, and false otherwise. Even though these lists can be useful in categorizing general text as hateful or aggressive, they are not well suited for classifying tweets as they are short and typically include modified words, URLs and emoticons. Overall, we find that bully and aggressive users have a minor bias towards using such words, but they are not significantly different from normal users' behavior.

4.3 Network-based features

The social network of Twitter plays a crucial role in diffusion of useful information and ideas, but also of negative opinions, rumors, and abusive language (e.g. [27, 40]). We study the association between aggressive or cyberbullying behavior and the position of users in the Twitter network of friends and followers. The network is comprised of about $1.2M$ users and $1.9M$ friend (i.e., someone who is followed by a user X) or follower (i.e., someone who follows a user X) edges, with 4.934 effective diameter, 0.0425 average clustering coefficient, and 24.95% and 99.99% nodes in the weakest and largest component, respectively. Users in such a network can have a varying degree of embeddedness with respect to friends or followers, reciprocity of connections, connectivity with different parts of the network, etc.



(a) Friends distribution. **(b)** Followers distribution.

(c) Reciprocity distribution. **(d)** Power difference distribution.

Figure 7: Network-based features: (a) Friends, and (b) Followers distribution, (c) Avg. distribution for Reciprocities, and (d) Power difference.

Popularity. The popularity of a user can be defined in different ways. For example, the number of friends or followers (out- or in-degree centrality), and the ratio of the two measures (since Twitter allows users to follow anyone without their approval, the ratio of followers to friends can quantify a user's popularity). These measures quantify the opportunity for a user to have a positive or negative impact in his ego-network in a direct way. Figures 7a and 7b indicate that bullies have fewer friends and followers than the other user categories, with normal users having the most friends.

Reciprocity. This metric quantifies the extent to which users reciprocate the follower connections they receive from other users. The average reciprocity in our network is 0.2. In Figure 7c we show that the user classes considered have different distributions, with the bully and aggressive users being more similar (i.e., higher number of reciprocities) than the normal or spam users. Reciprocity as a feature has also been used in [25], but in an interaction-based graph using likes in posts. Here, we investigate the fundamental reciprocity in Twitter friendship; the first to do this in the context of bullying.

Power Difference. A recent study [42] found that the emotional and behavioral state of victims depend on the power of their bullies, e.g., more negative emotional experiences were observed when more popular cyberbully users conducted the attack, and the high power difference with respect to status

in the network has been shown to be a significant characteristic of bullies [10]. Therefore, we consider a more elaborate feature: the power difference between a tweeter and his mentions. In fact, a further analysis of a user's mentioned users could reveal possible victims or bystanders of his aggressive or bullying behavior.

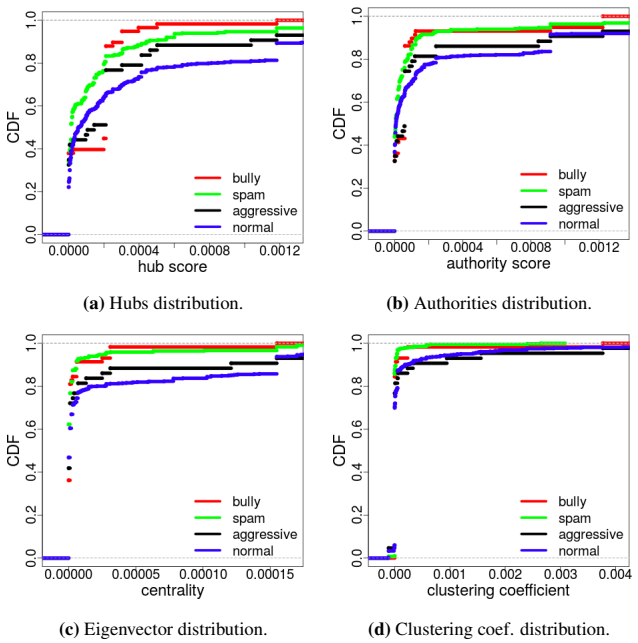
To this end, we compute the difference in power a user has with respect to the users he mentions in his posts in terms of their respective followers/friends ratio. Figure 7d shows the distribution of the power difference between a tweeter and his mentions (we note that the maximum power difference is 20, but we trim the plot for readability). The difference in power between the aggressive/bully and normal users is significant with D=0.22133 and D=0.31676, respectively.

Centrality Scores. We also investigate users' position in their network by considering more elaborate metrics that measure influence in their immediate and extended neighborhood, as well as connectivity. In particular, we study hub and authority centrality, as well as eigenvector and closeness centrality.

Hubs and Authority. A node's hub score is the sum of the authority score of the nodes that point to it, and authority shows how many different hubs a user is connected with [31].

Influence. Eigenvector centrality measures the influence of a user in his network, immediate or extended over multiple hops. Closeness centrality measures the extent to which a user is close to each other user in the network. To calculate the last four measures, we consider both the followers and friends relations of the users under examination in an undirected version of the network. Figures 8a, 8b, and 8c show the CDFs of the hubs (max value: 0.861), authorities (max value: 0.377), and eigenvector (max value: 0.0041) scores for the four classes of users. We observe that bully users tend to have lower values in their hub and authority scores which indicates they are not as popular within their networks. In terms of influence on their ego and extended network, they have behavior similar to spammers, while aggressors seem to have influence more similar to normal users. We omit the CDF of the closeness centrality measure because we are unable to reject the null hypothesis that the distributions are different.

Communities. Previous work [21] highlighted that bullies tend to experience social rejection from their environment and so they face difficulties in developing social relations. We examine the usefulness of this attribute and calculate the clustering coefficient measure, which shows a user's tendency to cluster with others of his network. Figure 8d plots the CDF of the clustering coefficient among the four behavioral classes. We observe that bully users, similarly to the spam ones, are less prone to create clusters in relation to aggressive and normal users. Finally, we compute communities using the Louvain method [2] which is suitable for identifying groups on large networks as it attempts to optimize the modularity measure (how densely connected the nodes within a cluster are) of a network by moving nodes from one cluster to another.



(a) Hubs distribution.

(b) Authorities distribution.

(c) Eigenvector distribution.

(d) Clustering coef. distribution.

Figure 8: Network-based features: Avg. distribution for (a) Hubs, (b) Authorities, (c) Eigenvectors, and (d) Clustering Coefficient.

Overall, we observe a few communities with a high number of nodes (especially the network core) resulting in a feature which differentiates bullies vs. normal users in a statistically significant fashion (D=0.206, p<0.05) but not aggressive vs. normal users (D=0.1166, p=0.6364).

5 Modeling Aggressors & Bullies

In this section, we present our efforts to model bullying and aggression behaviors on Twitter, using the features extracted and the labels provided by the crowdworkers. We examine the performance of several supervised classification algorithms using all and subsets of the labels provided.

5.1 Experimental Setup

We considered various machine learning algorithms, either probabilistic, tree-based, or ensemble classifiers (built on a set of classifiers whose individual decisions are then combined to classify data). Although we describe the classifiers we tested, due to space limitations, we only present the best results with respect to training time and performance, which were obtained with the Random Forest tree-based classifier. For all the experiments presented next we use the WEKA data mining toolkit and repeated (10 times) 10-fold cross validation [29], providing the relevant standard deviation (STD). It is important to note that we do not balance the data to better match a real-world deployment.

Features selection. Many of the features presented in Section 4 are found to be useful in discriminating between the classes. However, some are not useful and are excluded from the modeling analysis to avoid adding noise. Specifically,

we exclude the following features from our analysis: *user-based* - verified account, default profile image, statistics on sessions, *text-based* - average emotional scores, hate score, average word embedding, average curse score, and *network-based* - closeness centrality and Louvain modularity.

Tree-based classifiers. These classifiers are relatively fast compared to other classification models [43]. A tree classifier has three different types of nodes: (i) the *root* node, with no incoming edges, (ii) the *internal* nodes, with just one incoming edge and two or more outgoing edges, and (iii) the *leaf* node, with one incoming edge and no outgoing edges. The root and each internal node correspond to feature test conditions (in the simplest form, each test corresponds to a single feature) for separating data based on their characteristics, while the leaf nodes correspond to the available classes.

We experimented with various tree classifiers: J48, LADTree, LMT, NBTree, Random Forest (RF), and Functional Tree. We achieved the best performance with the RF classifier, which constructs a forest of decision trees with random subsets of features during the classification process. To build the model based on the RF, we proceed with the following parameter tuning: the number of trees to be generated is 10, and the maximum depth is unlimited.

Evaluation. For evaluation purposes, we examine standard machine learning performance metrics for the output model: *precision* (prec), *recall* (rec), and weighted area under the ROC curve (AUC), at the class level and overall average across classes. Also, the overall kappa (compares an observed accuracy with an expected accuracy, i.e., random chance), the root mean squared error (RMSE), which measures the differences among the values predicted by the model and the actually observed values, and finally the accuracy values are presented.

Experimentation phases. Two experimental setups are tested to assess the feasibility of detecting user behavior: (i) 4-classes classification, i.e., bully, aggressive, spam and normal users, and (ii) 3-classes classification, i.e., bully, aggressive and normal users. This setup examines the case where we filter out spam with a more elaborate technique and attempt to detect the bullies and aggressors from normal users.

5.2 Classification Results

Detecting offensive classes. Here, we examine whether it is possible to distinguish between bully, aggressive, spam and normal users. Table 2a overviews the results obtained with the RF classifier. In more detail, we observe that the classifier succeeds in detecting **43.2%** (STD=0.042) of the bully cases, which is quite satisfactory considering the small number of bully cases identified to begin with (only 4.5% of our dataset). In the aggressive case, we observe that recall is quite low, **11.8%** (STD=0.078). Based on the confusion matrix (omitted due to space limits), the misclassified cases mostly fall in either the normal or bullying classes, which is in alignment with the human annotations gathered during the crowdsourcing phase. Overall, the average precision is **71.6%**, and the recall is **73.32%**, while the accuracy equals to **73.45%** with 47.17% kappa and 30.86% RMSE.

	Prec.	**Rec.**	**AUC**
bully	0.411	0.432	0.893
(STD)	0.027	0.042	0.009
aggressive	0.295	0.118	0.793
(STD)	0.054	0.078	0.036
spammer	0.686	0.561	0.808
(STD)	0.008	0.010	0.002
normal	0.782	0.883	0.831
(STD.)	0.004	0.005	0.003
overall (avg.)	0.718	0.733	0.815
(STD)	0.005	0.004	0.031

(a) 4-classes classification

	Prec.	**Rec.**	**AUC**
bully	0.555	0.609	0.912
(STD)	0.018	0.029	0.009
aggressive	0.304	0.114	0.812
(STD)	0.039	0.012	0.015
normal	0.951	0.976	0.911
(STD)	0.018	0.029	0.009
overall (avg.)	0.899	0.917	0.907
(STD)	0.016	0.019	0.005

(b) 3-classes classification.

Table 2: Results on 4- and 3-classes classification.

Classifying after spam removal. In this experimental phase, we want to explore whether the distinction between bully/aggressive and normal users will be more evident after applying a more sophisticated spam removal process in the preprocessing step. To this end, we remove from our dataset all the spam related cases the annotators identified and re-run the classification process again with the RF classifier. Table 2b presents the results, where, as expected, for bully cases there is an important increase in both the precision (**14.4%**) and recall (**17.7%**). For aggressors, the precision and recall values are almost the same, indicating that further examination of their behavior is warranted in future work.

Overall, the average precision and recall of the RF model is **89.9%** and **91.7%**, respectively, while the accuracy is **91.08%** with 52.84% kappa value and 21.17% RMSE. Considering the AUC of **0.907**, we believe that with a more sophisticated spam detection applied on an incoming stream of tweets, our features and classification techniques can perform quite well at detecting bullies and aggressors and distinguishing them from the typical Twitter users.

Discussion on AUC. ROC curves are typically used to evaluate the performance of a machine learning algorithm [13] by testing the system on different points and getting pairs of true positive (i.e., recall) against false positive rates indicating the sensitivity of the used model. The resulting area under the ROC curve can be read as the probability of a classifier correctly ranking a random positive case higher than a random negative case. From Tables 2a, 2b we see that our model performs well. For both cases, we observe that in the aggressor case, even though the recall value is low, the AUC is quite high. This is due to the fact that the false positive rate is especially low, with 0.008 and 0.0135 for the 4-class and 3-class classification, respectively. We note that avoiding false positives is crucial to the successful deployment of any automated system aiming to deal with aggressive behavior.

Features evaluation. Table 3 shows the top 12 features for each experimental setup (based on information gain). Overall, in both experiments the most contributing features tend to be user- and network-based, which decribe how active and

Experiment	Feature (preserving order)
4-classes	#friends (11.43%), reciprocity (11.10%), #followers (10.84%) #followers/#friends (9.82%), interarrival (9.35%), #lists (9.11%) hubs (9.07%), #URLs (7.69%), #hashtags (7.02%) authority (6.33%), account age (4.77%), clustering coef. (3.47%)
3-classes	#followers/#friends (12.58%), #friends (12.56%), #followers (11.90%) interarrival (11.67%), reciprocity (11.01%), #lists (9.57%) hubs (8.41%), #hashtags (6.2%), #posts (6.01%) account age (4.13%), #URLs (3.73%), power difference (2.23%)

Table 3: Features evaluation.

	Prec.	Rec.	ROC
bully	1	0.667	0.833
aggressive	0.5	0.4	0.757
normal	0.931	0.971	0.82
overall (avg.)	0.909	0.913	0.817

Table 4: Classification on balanced data.

well-connected a user is with his network.

Balancing data. Based on [7], and similar to almost all classifiers, Random Forest suffers from appropriately handling extremely imbalanced training dataset (similar to our case) resulting in bias towards the majority classes. To address this issue, we over-sample (based on SMOTE [6], which creates synthetic instances of the minority class) and under-sample (a resampling technique without replacement) at the same time, as it has proven to result in better overall performance [6]. Here, we focus on the 3-class experimentation setup, i.e., without considering the spam user class. After randomly splitting the data into 90% for training and 10% for testing sets, we proceed with the balancing of the training set. The resulting data distribution is 349, 386, and 340 instances for the bully, aggressive, and normal classes, respectively. We note there is no resampling of the test set. Table 4 shows the classification results. After balancing the data, the classifier detects **66.7%** and **40%** of the bully and aggressive cases, respectively, while overall, the accuracy is 91.25%, with 59.65% kappa and 14.23% RMSE.

5.3 Twitter Reaction to Aggression

Recently, Twitter has received a lot of attention due to the increasing occurrence of harassment incidents [50]. While some shortcomings have been directly acknowledged by Twitter [49], they do act in some cases. To understand the impact of our findings, we make an estimate of Twitter's current effectiveness in dealing with harassment by looking at account statuses. Recall that Twitter accounts can be in one of three states: *active*, *deleted*, or *suspended*. Typically, Twitter suspends an account (temporarily or even permanently) if it has been hijacked/compromised, is considered spam/fake, or if it is *abusive* [52].

After our initial experimentation, we went back and checked the current status of all labeled users in our dataset. The status check was performed over two different time periods: at the end of November 2016 and February 2017. Tables 5a, 5b show the break down of account statuses for each label for the two time periods. From the more recent

	active	deleted	suspended
bully	67.24%	32.76%	0.00%
aggressive	65.12%	20.93%	13.95%
normal	86.53%	5.72%	7.75%

(a) Status check on Nov 2016.

	active	deleted	suspended
bully	62.07%	37.93%	0.00%
aggressive	55.81%	25.58%	18.60%
normal	85.01%	6.86%	8.13%

(b) Status check on Feb 2017.

Table 5: Distribution of users' behaviors in twitter statuses.

time period (February 2017), we observe that the majority of "bad" users in our dataset have suffered no consequences from Twitter: 55.81% of aggressive and 62.07% of cyberbullying accounts were still active. In particular, suspension (Twitter-taken action) and deletion (the user removing their account) statuses exhibit a stark contrast.

While 18.6% of aggressive users are suspended, *no* cyberbullying users are. Instead, cyberbullying users tend to delete their accounts proactively ($\sim 38\%$). Comparing the statuses of aggressive users between the two time periods, we see an increase (4.65%) in the percentage of those suspended. This is in alignment with Twitter's recent efforts to combat harassment cases, for instance, by preventing suspended users from creating new accounts [41], or temporarily limiting users for abusive behavior [48]. However, in both time periods, there are *no* suspended bully users. Again, bully users seem to delete their accounts, perhaps in an attempt to prevent suspension. Regardless of the reason, for these users we also observe a 5.17% increase in deleted accounts between the two time periods. The lack of suspension of bully users could be due to the fact that bullying often manifests in a hidden fashion, e.g., within seemingly innocuous criticisms, yet are repetitive (and thus harmful over time) in nature [33].

6 Related Work

We now review previous work on detecting offensive, abusive, aggressive, or bullying content/behavior on social media.

Detection. Over the past few years, several techniques have been proposed to measure and detect offensive or abusive content/behavior on platforms like Instagram [26], YouTube [9], 4chan [24], Yahoo Finance [15], and Yahoo Answers [28]. Chen et al. [9] use both textual and structural features (e.g., ratio of imperative sentences, adjective and adverbs as offensive words) to predict a user's aptitude in producing offensive content in Youtube comments, while Djuric et al. [15] rely on word embeddings to distinguish abusive comments on Yahoo Finance. Nobara et al. [38] perform hate speech detection on Yahoo Finance and News data, using supervised learning classification. Kayes et al. [28] find that users tend to flag abusive content posted on Yahoo Answers in an overwhelmingly correct way (as confirmed by human annotators). Also, some users significantly deviate from community norms, posting a large amount of content that is flagged as abusive. Through careful feature extraction, they also show it is possible to use machine learning methods to predict which users will be suspended.

Dinakar et al. [14] detect cyberbullying by decomposing it into detection of sensitive topics. They collect YouTube comments from controversial videos, use manual annotation to characterize them, and perform a bag-of-words driven text classification. Hee et al. [54] study linguistic characteristics in cyberbullying-related content extracted from Ask.fm, aiming to detect fine-grained types of cyberbullying types, such as threats and insults. Besides the victim and harasser, they also identify bystander-defenders and bystander-assistants, who support, respectively, the victim or the harasser. Finally, Hosseinmardi et al. [26] study images posted on Instagram and their associated comments to detect and distinguish between cyber-aggression and cyberbullying.

Attribute selection. A number of methods have been proposed to perform detection of harassment on social media. For instance, text features are often used to extract attributes that are in turn leveraged for classification. These include punctuations, URLs, part-of-speech, n-grams, Bag of Words (BoW), as well as lexical features relying on dictionaries of offensive words, and user-based features such as user's membership duration activity, number of friends/followers, etc. Different supervised approaches have been used for detection: [38] uses a regression model, whereas [12, 14, 54] rely on other methods like Naive Bayes, Support Vector Machines (SVM), and Decision Trees (J48). By contrast, Hosseinmardi et al. [25] use a graph-based approach based on likes and comments to build bipartite graphs and identify negative behavior. A similar graph-based approach is also used in [26].

Sentiment analysis of text can also contribute useful features in detecting offensive or abusive content. For instance, Nahar et al. [36] use sentiment scores of data collected from Kongregate (an online gaming site), Slashdot, and MySpace. They use a probabilistic sentiment analysis approach to distinguish between bullies and non-bullies, and rank the most influential users based on a predator-victim graph (built from exchanged messages). Finally, Xu et al. [57] rely on sentiment to identify victims on Twitter who pose high risk either to themselves or to others. Apart from using positive and negative sentiments, they also consider specific emotions, such as anger, embarrassment, and sadness.

Remarks. Overall, our work advances the state-of-art on cyberbullying and aggression detection by proposing a scalable methodology for large-scale analysis and extraction of text, user, and network based features on Twitter, which has not been studied in this context previously. Our novel methodology analyzes users' tweets, individually and in groups, and extracts appropriate features that connect user behavior with a tendency to be aggressive or bully. We examine the importance of such attributes, and further advance the state-of-art by focusing on new network-related attributes that further distinguish the specific user behaviors, specifically for Twitter. Finally, we discuss the effectiveness of our detection method by comparing results with the suspension and deletion of accounts as observed in the wild for users who, though aggressive, remain seemingly undetected.

7 Discussion & Conclusion

Although the digital revolution and the rise of social media enabled great advances in communication platforms and social interactions, wider proliferation of harmful behavior has also been facilitated. Unfortunately, effective tools for detecting harmful actions are scarce, as this type of behavior is often ambiguous in nature and/or exhibited via seemingly superficial comments and criticisms. Aiming to address this gap, this paper presented a novel system geared to automatically classify two kinds of harmful online behavior, cyber-aggression, and cyberbullying, focusing on the Twitter social network.

We relied on crowdsourced workers to label $1.5k$ users as normal, spammers, aggressive, or bullies, from a corpus of almost $10k$ tweets (distilled from a larger corpus of $1.6M$ tweets), using an efficient, streamlined labeling process. We investigated 30 features from 3 types of attributes – user, text, and network based – characterizing such behavior. We found that bully users are less popular (fewer followers/friends, lower hub, authority and eigenvector scores) and do not participate in many communities. Although they are not very active as per number of posts overall, when they do become active, they post more frequently than typical users, and do so with more hashtags, urls, etc. Interestingly, they also tend to have been long time Twitter users based on their account age.

Aggressive users show similar behavior to spammers in terms of number of followers, friends, and hub scores. Similar to bullies, they also do not post a lot of tweets, but exhibit a small response time between postings, and often use hashtags and URLs in their tweets. They also tend to have been on Twitter for a long time, like bullies. However, their posts seem to be more negative in sentiment than bullies or normal users. On the other hand, normal users are quite popular with respect to number of followers, friends, hubs, authorities. They participate in many topical lists, and use few hashtags and URLs. These observations are in line with the intuition that bully and aggressive users tend to attack, in rapid fashion, particular users or groups they target, and do so in short bursts, with not enough duration or content to be detected by Twitter's automated systems. In general, we find that aggressive users are more difficult to characterize and identify using a machine learning classifier than bullies, since sometimes they behave like bullies, but other times as normal or spam users.

We showed that our methodology for data analysis, labeling, and classification can scale up to millions of tweets, while our machine learning model built with a Random Forest classifier can distinguish between normal, aggressive, and cyberbullying users with high accuracy ($> 91\%$). While prior work almost exclusively focused on user- and text-based features (e.g., linguistics, sentiment, membership duration), we performed a thorough analysis of network-based features, and found them to be very useful, as they actually are the most effective for classifying aggressive user behavior (half

of the top-12 features in discriminatory power are network based). Whereas, text-based features, somewhat surprisingly, do contribute as much to the detection of aggression (with an exception of tweet characteristics, such as number of URLs, hashtags, and sentiment).

Finally, we discussed the effectiveness of our detection method by comparing prediction results of the users examined with the suspension and deletion of their accounts as observed in the wild. We found that users labeled as bullies tend not to have been suspended, but instead, take seemingly proactive measures and delete their accounts. On the other hand, aggressive users are suspended more often than bullies or normal users, which is in line with recent Twitter efforts to combat harassment cases by preventing suspended users from creating new accounts [41] or temporarily limiting users for abusive behavior [48].

Our work is a first step towards understanding and detecting aggressive and harassing users online. In the future, we aim to evaluate more sophisticated machine learning techniques and intervention strategies to move beyond detection and towards prevention.

Acknowledgements. This research has been fully funded by the European Commission as part of the ENCASE project (H2020-MSCA-RISE of the European Union under GA number 691025).

References

[1] J. Blackburn, R. Simha, N. Kourtellis, X. Zuo, M. Ripeanu, J. Skvoretz, and A. Iamnitchi. Branded with a scarlet "c": cheaters in a gaming social network. In *WWW*, pages 81–90, 2012.

[2] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. The louvain method for community detection in large networks. *Statistical Mechanics: Theory and Experiment*, 10:P10008, 2011.

[3] C. R. Center. Summary of our cyberbullying research (2004/16), November 2016. http://cyberbullying.org/summary-of-our-cyberbullying-research.

[4] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. Measuring #GamerGate: A Tale of Hate, Sexism, and Bullying. In *WWW CyberSafety Workshop*, 2017.

[5] D. Chatzakou, V. Koutsonikola, A. Vakali, and K. Kafetsios. Micro-blogging Content Analysis via Emotionally-Driven Clustering. In *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, 2002.

[7] C. Chen, A. Liaw, and L. Breiman. Using random forest to learn imbalanced data. *University of California, Berkeley*, pages 1–12, 2004.

[8] C. Chen, J. Zhang, X. Chen, Y. Xiang, and W. Zhou. 6 million spam tweets: A large ground truth for timely Twitter spam detection. In *IEEE ICC*, 2015.

[9] Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *PASSAT and SocialCom*, 2012.

[10] L. Corcoran, C. M. Guckin, and G. Prentice. Cyberbullying or cyber aggression?: A review of existing definitions of cyber-based peer-to-peer aggression. *Societies*, 5(2):245, 2015.

[11] CrowdFlower. crowdflower.com.

[12] M. Dadvar, D. Trieschnigg, and F. Jong. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Canadian Conference on Artificial Intelligence*, pages 275–281, 2014.

[13] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Machine learning*, pages 233–240. ACM, 2006.

[14] K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. *The Social Mobile Web*, 11:02, 2011.

[15] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate Speech Detection with Comment Embeddings. In *WWW*, 2015.

[16] Facts About Bullying. https://www.stopbullying.gov/news/media/facts/.

[17] J. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 1971.

[18] M. Giatsoglou, D. Chatzakou, N. Shah, C. Faloutsos, and A. Vakali. Retweeting Activity on Twitter: Signs of Deception. In *PAKDD (1)*, 2015.

[19] D. W. Grigg. Cyber-Aggression: Definition and Concept of Cyberbullying. *Australian Journal of Guidance and Counselling*, 20(2), 2010.

[20] D. W. Grigg. Cyber-aggression: Definition and concept of cyberbullying. *Australian Journal of Guidance and Counselling*, 20(02):143–156, 2010.

[21] L. D. Hanish, B. Kochenderfer-Ladd, R. A. Fabes, C. L. Martin, D. Denning, et al. Bullying among young children: The influence of peers and teachers. *Bullying in American schools: A social-ecological perspective on prevention and intervention*, pages 141–159, 2004.

[22] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[23] Hatebase database. ttps://www.hatebase.org/.

[24] G. E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM 2017*, 2017.

[25] H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and A. Ghasemianlangroodi. Towards understanding cyberbullying behavior in a semi-anonymous social network. In *IEEE/ACM ASONAM*, 2014.

[26] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. In *International Conference on Social Informatics*, 2015.

[27] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan. Epidemiological Modeling of News and Rumors on Twitter. In *SNAKDD*, 2013.

[28] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi. The Social World of Content Abusers in Community Question Answering. In *WWW*, 2015.

[29] J.-H. Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comput. Stat. Data Anal.*, 53(11):3735–3745, 2009.

[30] K. Kira and L. A. Rendell. A Practical Approach to Feature Selection. In *9th International Workshop on Machine Learning*, 1992.

[31] J. M. Kleinberg. Hubs, authorities, and communities. *ACM Comput. Surv.*, 31(4es), 1999.

[32] A. Massanari. #gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 2015.

[33] L. McMahon. Bullying and harassment in the workplace. *International Journal of Contemporary Hospitality Management*, 12(6):384–387, 2000.

[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[35] M. Miller. goo.gl/n1W6nt, Oct 2016.

[36] V. Nahar, S. Unankard, X. Li, and C. Pang. Sentiment Analysis for Effective Detection of Cyber Bullying. In *APWeb*, 2012.

[37] G. Navarro. A Guided Tour to Approximate String Matching. *ACM Comput. Surv.*, 33(1), 2001.

[38] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive Language Detection in Online User Content. In *WWW*, 2016.

[39] L. of Swear Words & Curse Words. http://www.noswearing.com/dictionary/.

[40] J. Pfeffer, T. Zorbach, and K. M. Carley. Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications*, 20(1-2), 2014.

[41] Pham, Sherisse. Twitter tries new measures in crackdown on harassment. CNNtech, February 2017. goo.gl/nMi4ZQ.

[42] S. Pieschl, T. Porsch, T. Kahl, and R. Klockenbusch. Relevant dimensions of cyberbullying - results from two experimental studies. *Journal of Applied Developmental Psychology*, 34(5):241 – 252, 2013.

[43] J. Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1), 1986.

[44] Twitter trolls are now abusing the company's bottom line. goo.gl/SryS3k, 2016.

[45] SentiStrength. http://sentistrength.wlv.ac.uk/.

[46] P. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett. Cyberbullying: Its nature and impact in secondary school pupils. In *Child Psychol. Psychiatr*, 2008.

[47] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *ACSAC*, 2010.

[48] A. Sulleyman. Twitter temporarily limiting users for abusive behaviour. Independent, February 2017. goo.gl/yfJrZn.

[49] The Guardian. Twitter CEO: We suck at dealing with trolls and abuse. goo.gl/6CxnwP, 2015.

[50] The Guardian. Did trolls cost Twitter 3.5bn and its sale? goo.gl/2IdA5W, 2016.

[51] R. S. Tokunaga. Review: Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization. *Comput. Hum. Behav.*, 26(3), 2010.

[52] Twitter. About suspended accounts. https://support.twitter.com/articles/15790.

[53] UMICH SI650 - Sentiment Classification. https://inclass.kaggle.com/c/si650winter11, Apr 2011.

[54] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste. Automatic detection and prevention of cyberbullying. In *Human and Social Analytics*, pages 13–18, 2015.

[55] A. Viera and J. Garrett. Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5):360–363, 5 2005.

[56] A. H. Wang. Don't follow me: Spam detection in Twitter. In *SECRYPT*, 2010.

[57] J.-M. Xu, X. Zhu, and A. Bellmore. Fast Learning for Sentiment Analysis on Bullying. In *WISDOM*, 2012.