Weakly Supervised Cyberbullying Detection with Participant-Vocabulary Consistency

Elaheh Raisi
Department of Computer Science
Virginia Tech
Email: elaheh@vt.edu

Bert Huang
Department of Computer Science
Virginia Tech
Email: bhuang@vt.edu

***Abstract*—Online harassment and cyberbullying are becoming serious social health threats damaging people's lives. This phenomenon is creating a need for automated, data-driven techniques for analyzing and detecting such detrimental online behaviors. We propose a weakly supervised machine learning method for simultaneously inferring user roles in harassment-based bullying and new vocabulary indicators of bullying. The learning algorithm considers social structure and infers which users tend to bully and which tend to be victimized. To address the elusive nature of cyberbullying using minimal effort and cost, the learning algorithm only requires weak supervision. The weak supervision is in the form of expert-provided small seed of bullying indicators, and the algorithm uses a large, unlabeled corpus of social media interactions to extract bullying roles of users and additional vocabulary indicators of bullying. The model estimates whether each social interaction is bullying based on who participates and based on what language is used, and it tries to maximize the agreement between these estimates, i.e., participant-vocabulary consistency (PVC). To evaluate PVC, we perform extensive quantitative and qualitative experiments on three social media datasets: Twitter, Ask.fm, and Instagram. We illustrate the strengths and weaknesses of the model by analyzing the identified conversations and key phrases by PVC. In addition, we demonstrate the distributions of bully and victim scores to examine the relationship between the tendencies of users to bully or to be victimized. We also perform fairness evaluation to analyze the potential for automated detection to be biased against particular groups.**

I. Introduction

A growing portion of human communication is occurring over Internet services, and advances in mobile and networked technology have amplified individuals' abilities to connect and stay connected to each other. Moreover, the digital nature of these services enables them to measure and record unprecedented amounts of data about social interactions. Unfortunately, the amplification of social connectivity also includes the amplification of negative aspects of society, leading to significant phenomena such as online harassment, cyberbullying, hate speech, and online trolling [1]–[6]. StopBullying.gov defines cyberbullying as "bullying that takes place using electronic technology[, including] devices and equipment such as cell phones, computers, and tablets as well as communication tools including social media sites, text messages, chat, and websites." Three criteria define traditional bullying: (a) intent to cause harm, (b) repetition of the behavior over time, and (c) an imbalance of power between the victim(s) and bully(ies) [7]–[9]. In seeking formal definitions for cyberbullying, the central question has been whether the same criteria can be used [10]–[13]. Aggression and repetition are the two key elements of bullying that translate to the online setting. However, power imbalance is nontrivial to characterize online. In traditional bullying, power imbalance is often straightforward, such as a difference in physical strength. In online settings, various forms of power, such as anonymity, the constant possibility of threats, and the potential for a large audience, can create power imbalances in cyberbullying [14]. These factors make the design of automated cyberbullying detection a challenge that can benefit from machine learning.

According to *stopbullying.gov*, there are various forms of cyberbullying, including but not limited to harassment, rumor spreading, and posting of embarrassing images. In this study, we focus on harassment, in which harassers (bullies) send toxic and harmful communications to victims. We present an automated, data-driven method for identification of harassment. Our approach uses machine learning with weak supervision, significantly alleviating the need for human experts to perform tedious data annotation.

Analysis of online harassment requires multifaceted understanding of language and social structures. The complexities underlying these behaviors make automatic detection difficult for static computational approaches. For example, keyword searches or sentiment analyses are insufficient to identify instances of harassment, as existing sentiment analysis tools often use fixed keyword lists [15]. In contrast, fully supervised machine learning enables models to be adaptive. However, to train a supervised machine learning method, labeled input data is necessary. Data annotation is a costly and time-demanding process. High-quality hand-labeled data is a key bottleneck in machine learning. Therefore, many researchers have been developing weakly supervised algorithms in which only a limited amount of data is labeled. The intuition behind weak supervision is that the learning algorithm should be able to find patterns in the unlabeled data to integrate with the weak supervision. We use this weak supervision paradigm to significantly alleviate the need for human experts to perform tedious data annotation. Our weak supervision is in the form of expert-provided key phrases that are highly indicative of bullying. For example, various swear words and slurs are common indicators of bullying. The algorithms then infer

unknown data values from these expert annotations to find instances of bullying.

The algorithm we present here learns a relational model by using the structure of the communication network. The relational model is trained in a weakly supervised manner, where human experts only need to provide high-fidelity annotations in the form of key phrases that are highly indicative of harassment. The algorithm then extrapolates from these expert annotations—by searching for patterns of victimization in an unlabeled social interaction network—to find other likely key-phrase indicators and specific instances of bullying.

We refer to the proposed method as the *participant-vocabulary consistency* (PVC) model. The algorithm seeks a consistent parameter setting for all users and key phrases in the data that characterizes the tendency of each user to harass or to be harassed and the tendency of a key phrase to be indicative of harassment. The learning algorithm optimizes the parameters to minimize their disagreement with the training data, which takes the form of a directed message network, with each message acting as an edge decorated by its text content. PVC thus fits the parameters to patterns of language use and social interaction structure.

An alarming amount of harassment occurs in public-facing social media, such as public comments on blogs and media-sharing sites. We will use this type of data as a testbed for our algorithms. According to a survey by *ditchthelabel.org* [16], the five sites with the highest concentration of cyberbullying are Facebook, YouTube, Twitter, Ask.fm, and Instagram. We evaluate participant-vocabulary consistency on social media data from three of these sources: Twitter, Ask.fm, and Instagram. We use a human-curated list of key phrases highly indicative of bullying as the weak supervision, and we test how well participant-vocabulary consistency identifies examples of bullying interactions and new bullying indicators.

We conduct wide range of quantitative and qualitative experiments to examine how well PVC identifies examples of bullying interactions and new bullying indicators. In our quantitative evaluation, we use post-hoc human annotation to measure how well PVC fits human opinions about bullying. In our qualitative analysis, we group the identified conversations by PVC into three categories: 1) true positives that other baselines were not be able to detect, 2) true positives not containing very obvious offensive languages, and 3) false positives. We inspect the false positives and notice there are four different types: (1) users talking about other people, not addressing each other in their messages, (2) joking conversations, (3) users talking about some bullying-related topics, (4) conversations with no language indicative of bullying. We also analyze another group of false positives we call *unfair false positives*. Fairness is an important topic when considering any online automated harassment detection. We measure the sensitivity of PVC to language describing particular social groups, such as those defined by race, gender, sexual orientation, and religion. In another set of qualitative evaluations, we show the relationship between the learned user's bully and victim scores in heatmap and scatter plots. We also provide a summary statistics about bullies and victims such as their average in-degree and out-degree. In addition, we showed a few small sub-networks of identified bullies and victims to see how differently they are distributed around each other.

The main contributions of this paper are as follows: We present the participant-vocabulary consistency model, a weakly supervised approach for simultaneously learning the roles of social media users in the harassment form of cyberbullying and the tendency of language indicators to be used in such cyberbullying. We demonstrate that PVC can discover examples of apparent bullying as well as new bullying indicators, in part because the learning process of PVC considers the structure of the communication network. We evaluate PVC on a variety of social media data sets with both quantitative and qualitative analyses. This method is the first specialized algorithm for cyberbullying detection that allows weak supervision and uses social structure to simultaneously make dependent, collective estimates of user roles in cyberbullying and new cyberbullying language indicators.

II. Background and Related Work

In this section, we briefly summarize the related work that we build upon. The two main bodies of research that support our contribution are (1) emerging research investigating online harassment and cyberbullying, and (2) research developing automated methods for vocabulary discovery.

A variety of methods have been proposed for cyberbullying detection. These methods mostly approach the problem by treating it as a classification task, where messages are independently classified as bullying or not. Many of the research contributions in this space involve the specialized design of language features for supervised learning. Such feature design is complementary to our approach and could be seamlessly incorporated into our framework. Many contributions consider specially designed features based on known topics used in bullying [17]–[20]. Others use sentiment features [21], features learned by topic models [22], vulgar language expansion using string similarity [23], features based on association rule techniques [24], and static, social structure features [25]–[27]. Some researchers used probabilistic fusion methods to combine social and text features together as the input of classifier [28]. Researchers have applied machine learning methods to better understand social-psychological issues surrounding the idea of bullying [29]. By extracting tweets containing the word "bully," they collect a data set of people talking about their experiences with bullying. They also investigate different forms of bullying and why people post about bullying. Additionally, some studies have extensively involved firsthand accounts of young persons, yielding insights on new features for bullying detection and strategies for mitigation [30].

A group of researchers studied the behavior of bullies and which features distinguish them from regular users by extracting text, user, and network-based attributes [31]. Similarly, some examined the properties of cyber-aggressors, their posts, and their difference from other users in the content of the *Gamergate controversy* [32], [33]. Other researchers analyzed

the online setting of cyberbullying detection by extracting a small set of social network structure features that are the most important to cyberbullying to improve time and accuracy [34], [35]. Some separate tweets into three categories: those containing hate speech, only offensive language, and those with neither. They trained a supervised three-class classifiers using language features [36]. In addition, some research considering firsthand accounts of young persons shed lights on new features for bullying detection and strategies for mitigation [30].

Hosseinmardi et al. conducted several studies analyzing cyberbullying on Ask.fm and Instagram. They studied negative user behavior in the Ask.fm social network, finding that properties of the interaction graph—such as in-degree and out-degree—are strongly related to negative or positive user behaviors [37].

They compared users across Instagram and Ask.fm to see how negative user behavior varies across different venues. Based on their experiments, Ask.fm users show more negativity than Instagram users, and anonymity on Ask.fm tends to foster more negativity [38]. They also studied the detection of cyberbullying incidents over images on Instagram, focusing on the distinction between cyberbullying and cyber-aggression [39], noting that bullying occurs over multiple interactions with particular social structures.

Related research on data-driven methods for analysis and detection of cyberviolence in general includes detection of hate speech [40]–[42], online predation [43], and the analysis of gang activity on social media [44], among many other emerging projects.

Our proposed method simultaneously learns new language indicators of bullying while estimating users' roles in bullying behavior. Learning new language indicators is related to the task of query expansion in information retrieval [45]. Query expansion aims to suggest a set of related keywords for user-provided queries. Massoudi et al. [46] use temporal information as well as co-occurrence to score the related terms to expand the query. Lavrenko et al. [47] introduce a relevance-based approach for query expansion by creating a statistical language model for the query. This commonly-used approach estimates the probabilities of words in the relevant class using the query alone. Mahendiran et al. [48] propose a method based on probabilistic soft logic to grow a vocabulary using multiple indicators (social network, demographics, and time). They apply their method to expand the political vocabulary of presidential elections.

Preliminary results from the research here appeared in a short, non-archival paper for a workshop [49]. A conference paper [50] presented extended and complete description, analysis, and results from the study. This article extends the conference paper results with more thorough analysis of the performance and behavior of the proposed method.

III. Participant-Vocabulary Consistency

Our weakly supervised approach is built on the idea that it should be inexpensive for human experts to provide weak indicators of some forms of bullying, specifically vocabulary commonly used in bullying messages. The algorithm extrapolates from the weak indicators to find possible instances of bullying in the data. Then, considering the discovered users who tend to be involved in bullying, the algorithm finds new vocabulary that is commonly used by these suspected bullies and victims. This feedback loop iterates until the algorithm converges on a consistent set of scores for how much the model considers each user to be a bully or a victim, and a set of scores for how much each vocabulary key-phrase is an indicator of bullying. The idea is that these vocabulary scores will expand upon the language provided in the weak supervision to related terminology, as well as to language used in different types of bullying behavior. The algorithm considers the entire network of communication, propagating its estimates of bullying roles through the messaging structure and the language used in each message, leading to a joint, collective estimation of bullying roles across the network.

We use a general data representation that is applicable to a wide variety of social media platforms. To formalize the observable data from such platforms, we first consider a set of users U and a set of messages M. Each message $m \in M$ is sent from user $s(m)$ to user $r(m)$. I.e., the lookup functions s and r return the sender and receiver, respectively, of their input message. Each message m is described by a set of feature occurrences $f(m) := \{x_k, \ldots, x_\ell\}$. Each feature represents the existence of some descriptor in the message. In our experiments and in many natural instantiations of this model, these descriptors represent the presence of n-grams in the message text, so we will interchangeably refer to them as vocabulary features.

For example, if m is a Twitter message from user @alice with the text "@bob hello world", then

$$s(m) = \text{@alice}, \quad r(m) = \text{@bob}$$
$$f(m) = \{\text{hello}, \text{world}, \text{hello world}\}.$$

In this representation, a data set can contain multiple messages from or to any user, and multiple messages involving the same pair of users. E.g., @alice may send more messages to @bob, and they may contain completely different features.

To model cyberbullying roles, we attribute each user u_i with a bully score b_i and a victim score v_i. The bully score encodes how much our model believes a user has a tendency to bully others, and the victim score encodes how much our model believes a user has a tendency to be bullied. We attribute to each feature x_k a bullying-vocabulary score w_k, which encodes how much the presence of that feature indicates a bullying interaction.

For each message sent from user u_i to user u_j, we use an additive *participant score* combining the sender's bully score and the receiver's victim score $(b_i + v_j)$. The more the model believes u_i is a bully and u_j is a victim, the more it should believe this message is an instance of bullying. To predict the bullying score for each interaction, we combine the total

average word score of the message with the participant score

$$\underbrace{\left(b_{s(m)} + v_{r(m)}\right)}_{\text{participant score}} + \underbrace{\frac{1}{|f(m)|}\sum_{k \in f(m)} w_k}_{\text{vocabulary score}} . \quad (1)$$

We then define a regularized objective function that penalizes disagreement between the social bullying score and each of the message's bullying-vocabulary scores:

$$J(\mathbf{b}, \mathbf{v}, \mathbf{w}) = \tfrac{\lambda}{2}\left(||\mathbf{b}||^2 + ||\mathbf{v}||^2 + ||\mathbf{w}||^2\right) + \tfrac{1}{2}\sum_{m \in M}\left(\sum_{k \in f(m)}\left(b_{s(m)} + v_{r(m)} - w_k\right)^2\right). \quad (2)$$

The learning algorithm seeks settings for the $\mathbf{b}$, $\mathbf{v}$, and $\mathbf{w}$ vectors that are consistent with the observed social data and initial *seed features*. We have used a joint regularization parameter for the word scores, bullying scores, and victim scores, but it is easy to use separate parameters for each parameter vector. We found in our experiments that the learner is not very sensitive to these hyperparameters, so we use a single parameter λ for simplicity. We constrain the seed features to have a high score and minimize Eq. (2), i.e.,

$$\min_{\mathbf{b},\mathbf{v},\mathbf{w}} \; J(\mathbf{b}, \mathbf{v}, \mathbf{w}; \lambda) \text{ s.t. } w_k = 1.0, \; \forall k : x_k \in S, \quad (3)$$

where S is the set of seed words. By solving for these parameters, we optimize the consistency of scores computed based on the participants in each social interaction as well as the vocabulary used in each interaction. Thus, we refer to this model as the participant-vocabulary consistency model.

A. Alternating Least Squares

The objective function in Eq. (2) is not jointly convex, but it is convex when optimizing each parameter vector in isolation. In fact, the form of the objective yields an efficient, closed-form minimization for each vector. The minimum for each parameter vector considering the others constant can be found by solving for their zero-gradient conditions. The solution for optimizing with respect to $\mathbf{b}$ is

$$\arg\min_{b_i} J = \frac{\sum_{m \in M | s(m) = i}\left(\sum_{k \in f(m)} w_k - |f(m)| v_{r(m)}\right)}{\lambda + \sum_{m \in M | s(m) = i} |f(m)|}, \quad (4)$$

where the set $\{m \in M | s(m) = i\}$ is the set of messages that are sent by user i, and $|f(m)|$ is the number of n-grams in the message m. The update for the victim scores $\mathbf{v}$ is analogously

$$\arg\min_{v_j} J = \frac{\sum_{m \in M | r(m) = j}\left(\sum_{k \in f(m)} w_k - |f(m)| b_i\right)}{\lambda + \sum_{m \in M | r(m) = j} |f(m)|}, \quad (5)$$

where the set $\{m \in M | r(m) = j\}$ is the set of messages sent to user j. Finally, the update for the $\mathbf{w}$ vector is

$$\arg\min_{w_k} J = \frac{\sum_{m \in M | k \in f(m)}\left(b_{r(m)} + v_{s(m)}\right)}{\lambda + |\{m \in M | k \in f(m)\}|}, \quad (6)$$

where the set $\{m \in M | k \in f(m)\}$ is the set of messages that contain the kth feature or n-gram.

Each of these minimizations solves a least-squares problem, and when the parameters are updated according to these formulas, the objective is guaranteed to decrease if the current parameters are not a local minimum. Since each formula of the $\mathbf{b}$, $\mathbf{v}$, and $\mathbf{w}$ vectors does not depend on other entries within the same vector, each full vector can be updated in parallel. Thus, we use an alternating least-squares optimization procedure, summarized in Algorithm 1, which iteratively updates each of these vectors until convergence.

Algorithm 1 Participant-Vocabulary Consistency using Alternating Least Squares

procedure PVC(b, v, w, λ)
  Initialize $\mathbf{b}$, $\mathbf{v}$, and $\mathbf{w}$ to default values (e.g., 0.1).
  while not converged **do**
    $\mathbf{b} = \left[\arg\min_{b_i} J\right]_{i=1}^{n}$ ▷ update $\mathbf{b}$ using Eq. (4)
    $\mathbf{v} = \left[\arg\min_{v_i} J\right]_{i=1}^{n}$ ▷ update $\mathbf{v}$ using Eq. (5)
    $\mathbf{w} = \left[\arg\min_{w_k} J\right]_{k=1}^{|\mathbf{w}|}$ ▷ update $\mathbf{w}$ using Eq. (6)
  return $(\mathbf{b}, \mathbf{v}, \mathbf{w})$ ▷ return the final bully, victim score of users and the score of n-grams

Algorithm 1 outputs the bully and victim score of all the users and the bullying-vocabulary score of all n-grams. Let $|M|$ be the total number of messages and $|W|$ be the total number of n-grams. The time complexity of each alternating least-squares update for the bully score, victim score, and word score is $O(|M| \cdot |W|)$. No extra space is needed beyond the storage of these vectors and the raw data. Moreover, sparse matrices can be used to perform the indexing necessary to compute these updates efficiently and conveniently, at no extra cost in storage, and the algorithm can be easily implemented using high-level, optimized sparse matrix libraries. E.g., we use `scipy.sparse` for our implementation.

IV. Experiments

We apply participant-vocabulary consistency to detect harassment-based bullying in three social media data sets, and we measure the success of weakly supervised methods for detecting examples of cyberbullying and discovering new bullying indicators. We collect a dictionary of offensive language listed on NoSwearing.com [51]. This dictionary contains 3,461 offensive unigrams and bigrams. We then compare human annotations against PVC and baseline methods for detecting cyberbullying using the provided weak supervision. We also compare each method's ability to discover new bullying vocabulary, using human annotation as well as cross-validation

tests. Finally, we perform qualitative analysis of the behavior of PVC and the baselines on each data set.

To set the PVC regularization parameter λ, we use three-fold cross-validation; i.e., we randomly partition the set of collected offensive words into three complementary subsets, using each as a seed set in every run. We do not split the user or message data since they are never directly supervised. For each fold, we use one third of these terms to form a seed set for training. We refer to the remaining held-out set of offensive words in the dictionary as *target words*. (The target words include bigrams as well, but for convenience we refer to them as target words throughout.) We measure the average area under the receiver order characteristic curve (AUC) for target-word recovery with different values of λ from 0.001 to 20.0. The best value of λ should yield the largest AUC. The average AUC we obtain using these values of λ for three random splits of our Twitter data (described below) ranged between 0.905 and 0.928, showing minor sensitivity to this parameter. Based on these results, we set $\lambda = 8$ in our experiments, and for consistency with this parameter search, we run our experiments using one of these random splits of the seed set. Thus, we begin with just over a thousand seed phrases, randomly sampled from our full list.

A. Data Processing

Ask.fm, Instagram, and Twitter are reported to be key social networking venues where users experience cyberbullying [16], [52], [53]. Our experiments use data from these sources.

We collected data from **Twitter**'s public API. Our process for collecting our Twitter data set was as follows: (1) Using our collected offensive-language dictionary, we extracted tweets containing these words posted between November 1, 2015, and December 14, 2015. For every curse word, we extracted 700 tweets. (2) Since the extracted tweets in the previous step were often part of a conversation, we extracted all the conversations and reply chains these tweets were part of. (3) To avoid having a skewed data set, we applied snowball sampling to expand the size of the data set, gathering tweets in a wide range of topics. To do so, we randomly selected 1,000 users; then for 50 of their followers, we extracted their most recent 200 tweets. We continued expanding to followers of followers in a depth-10 breadth-first search. Many users had small follower counts, so we needed a depth of 10 to obtain a reasonable number of these background tweets.

We filtered the data to include only public, directed messages, i.e., @-messages. We then removed all retweets and duplicate tweets. After this preprocessing, our Twitter data contains 180,355 users and 296,308 tweets. Once we obtained the conversation structure, we then further processed the message text, removing emojis, mentions, and all types of URLs, punctuation, and stop words.

We used the **Ask.fm** data set collected by Hosseinmardi et al. [38]. On Ask.fm, users can post questions on public profiles of other users, anonymously or with their identities revealed. The original data collection used snowball sampling, collecting user profile information and a complete list of answered questions. Since our model calculates the bully and victim scores for every user, it does not readily handle anonymous users, so we removed all the question-answer pairs where the identity of the question poster is hidden. Furthermore, we removed question-answer pairs where users only post the word "thanks" and nothing else, because this was extremely common and not informative to our study. Our filtered data set contains 260,800 users and 2,863,801 question-answer pairs. We cleaned the data by performing the same preprocessing steps as with Twitter, as well as some additional data cleaning such as removal of HTML tags.

We used the **Instagram** data set collected by Hosseinmardi et al. [54], who identified Instagram user IDs using snowball sampling starting from a random seed node. For each user, they collected all the media the user shared, users who commented on the media, and the comments posted on the media. Our Instagram data contains 3,829,756 users and 9,828,760 messages.

B. Baselines

Few alternate approaches have been established to handle weakly supervised learning for cyberbullying detection. The most straightforward baseline is to directly use the weak supervision to detect bullying, by treating the seed key-phrases as a search query.

To measure the benefits of PVC's learning of user roles, we compare against a method that extracts participant and vocabulary scores using only the seed query. For each user, we compute a bullying score as the fraction of outgoing messages that contain at least one seed term over all messages sent by that user and a victim score as the fraction of all incoming messages that contain at least one seed term over all messages received by that user. For each message, the participant score is the summation of the sender's bullying score and the receiver's victim score. We also assign each message a vocabulary score computed as the fraction of seed terms in the message. As in PVC, we sum the participant and vocabulary scores to compute the score of each message. We refer to this method in our results as the *naive participant* method.

We also compare against existing approaches that expand the seed query. This expansion is important for improving the recall of the detections, since the seed set will not include new slang or may exclude indicators for forms of bullying the expert annotators neglected. The key challenge in what is essentially the expansion of a search query is maintaining a high precision as the recall is increased. We compare PVC to two standard heuristic approaches for growing a vocabulary from an initial seed query. We briefly describe each below.

Co-occurrence (CO) returns any word (or n-gram) that occurs in the same message as any of the seed words. It extracts all messages containing any of the seed words and considers any other words in these messages to be relevant key-phrases. All other words receive a score of 0. We should expect co-occurrence to predict a huge number of words, obtaining high recall on the target words but at the cost of collecting large amounts of irrelevant words.

Dynamic query expansion (DQE) is a more robust variation of co-occurrence that iteratively grows a query dictionary by considering both co-occurrence and frequency [55]. We use a variation based on phrase relevance. Starting from the seed query, DQE first extracts the messages containing seed phrases; then for every term in the extracted messages, it computes a relevance score (based on [47]) as the rate of occurrence in relevant messages: $\text{relevance}(w_i, d, D) = |d \in D : w_i \in d|/|D|$, where $|D|$ indicates the number of documents with at least one seed term. Next, DQE picks k of the highest-scoring keywords for the second iteration. It continues this process until the set of keywords and their relevance scores become stable. Because DQE seeks more precise vocabulary expansion by limiting the added words with a parameter k, we expect it to be a more precise baseline, but in the extreme, it will behave similarly to the co-occurrence baseline. In our experiments, we use $k =$ 4,000, which provides relatively high precision at the cost of relatively low recall.

C. Human Annotation Comparisons

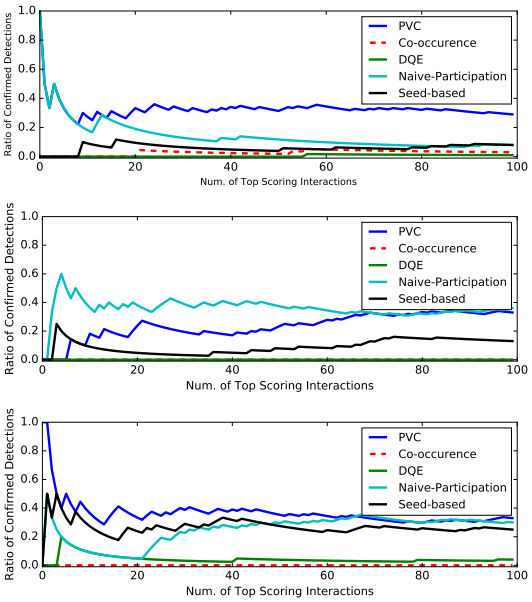

Fig. 1: Precision@k for bullying interactions on Ask.fm (top), Instagram (middle), and Twitter (bottom).

The first form of evaluation we perform uses post-hoc human annotation to rate how well the outputs of the algorithms agree with annotator opinions about bullying. We enlisted crowdsourcing workers from Amazon Mechanical Turk, restricting the users to Mechanical Turk Masters located in the United States. We asked the annotators to evaluate the outputs of the three approaches from two perspectives: the discovery of cyberbullying relationships and the discovery of additional language indicators. First, we extracted the 100 directed user pairs most indicated to be bullying by each method. For the PVC and naive-participant methods, we averaged the combined participant and vocabulary scores, as in Eq. (1), of all messages from one user to the other. For the dictionary-based baselines, we scored each user pair by the concentration of detected bullying words in messages between the pair. Then we collected all interactions between each user pair in our data. We showed the annotators the anonymized conversations and asked them, "Do you think either user 1 or user 2 is harassing the other?" The annotators indicated either "yes," "no," or "uncertain." We collected five annotations per conversation.

Second, we asked the annotators to rate the 1,000 highest-scoring terms from each method, excluding the seed words. These represent newly discovered vocabulary the methods believe to be indicators of harassment. For co-occurrence, we randomly selected 1,000 co-occurring terms among the total co-occurring phrases. We asked the annotators, "Do you think use of this word or phrase is a potential indicator of harassment?" We collected three annotations per key-phrase.

In Fig. 1, we plot the precision@k of the top 100 interactions for each data set and each method. The precision@k is the proportion of the top k interactions returned by each method that the majority of annotators agreed seemed like bullying. For each of the five annotators, we score a positive response as +1, a negative response as -1, and an uncertain response as 0. We sum these annotation scores for each interaction, and we consider the interaction to be harassment if the score is greater than or equal to 3. In the Ask.fm data, PVC significantly dominates the other methods for all thresholds. On the Twitter data, PVC is better than baselines until approximately interaction 70, when it gets close to the performance of the naive-participant baseline. In the Instagram data, PVC is below the precision of the naive-participation score until around interaction 65, but after that it improves to be the same as naive-participant. Co-occurrence, while simple to implement, appears to expand the dictionary too liberally, leading to very poor precision. DQE expands the dictionary more selectively, but still leads to worse precision than using the seed set alone.

In Fig. 2, we plot the precision@k for indicators that the majority of annotators agreed were indicators of bullying. On all three data sets, PVC detects bullying words significantly more frequently than the two baselines, again demonstrating the importance of the model's simultaneous consideration of the entire communication network.

It is useful to note that the performance of the algorithm is directly affected by the quality and quantity of seed words. A better hand-picked seed set will result in higher precision as our model is founded based on this set. If the number of indicator words in the seed set increases, we expect increased recall but decreased precision. Adding a poor indicator word to the seed set may result in reducing both precision and recall, because the algorithm may identify non-bullying conversations as bullying, and consequently increasing the false positive rate. Moreover, by filtering the seed set, it is possible to focus PVC on particular topics of bullying.

D. Qualitative Analysis

We analyzed the 1,000 highest-scoring, non-seed terms produced by PVC, DQE, and co-occurrence and categorized

TABLE I: Color-coded bullying bigrams detected in Ask.fm data by PVC and baselines. Terms are categorized according to the aggregate score of annotations. "Bullying" (2 or greater), "Likely Bullying" (1), "Uncertain" (0), and "Not Bullying" (negative) bigrams are shown in red, orange, gray, and blue, respectively.

Method	Detected Bullying Words Color-Coded by Annotation: Bullying, Likely Bullying, Uncertain, Not Bullying.
PVC	oreo nice, massive bear, bear c*ck, f*cking anus, ure lucky, f*g f*g, d*ck b*tch, ew creep, f*cking bothering, rupture, f*cking p*ssy, support gay, house f*ggot, family idiot, b*tch b*tch, p*ssy b*tch, loveeeeeee d*ck, f*cking c*nt, penis penis, gross bye, taste nasty, f*cking f*cking, dumb hoe, yellow attractive, b*tch p*ssy, songcried, songcried lika, lika b*tch, b*tch stupid, um b*tch, f*cking obv, nice butt, rate f*g, f*cking stupid, juicy red, soft juicy, f*cking d*ck, cm punk, d*ck p*ssy, stupid f*cking, gay bestfriend, eat d*ck, ihy f*g, gay gay, b*tch f*cking, dumb wh*re, s*ck c*ck, gay bi, fight p*ssy, stupid hoe
DQE	lol, haha, love, tbh, hey, yeah, good, kik, ya, talk, nice, pretty, idk, text, hahaha, rate, omg, xd, follow, xx, ty, funny, cute, people, cool, f*ck, best, likes, ily, sh*t, beautiful, perfect, girl, time, going, hot, truth, friends, lmao, answers, hate, ik, thoughts, friend, day, gonna, ma, gorgeous, anon, school
CO	bby, ana, cutie, ikr, ja, thnx, mee, profile, bs, feature, plz, age, add, pls, wat, ka, favourite, s*cks, si, pap, promise, mooi, hii, noo, nu, blue, ben, ook, mn, merci, meh, men, okk, okayy, hbu, zelf, du, dp rate, mooie, fansign, english, best feature, basketball, meisje, yesss, tyy, shu, een, return, follow follow

TABLE II: Color-coded bullying bigrams detected in Instagram data by PVC and baselines

Method	Detected Bullying Words Color-Coded by Annotation: Bullying, Likely Bullying, Uncertain, Not Bullying.
PVC	b*tch yas, yas b*tch, b*tch reported, *ss *ss, treated ariana, kitty warm, warm kitty, chicken butt, happy sl*t, jenette s*cking, kitty sleepy, follower thirsty, ariana hope, *ss b*tch, tart deco, sleepy kitty, hatejennette, *ss hoe, b*tch b*tch, sl*t hatejennette, pays leads, deco, happy kitty, fur happy, black yellow, bad *ss, bad b*tch, yellow black, pur pur, kitty pur, black black, d*ck b*tch, boss *ss, b*tch s*ck, soft kitty, nasty *ss, kitty purr, stupid *ss, *sss *ss, stupid b*tch, puff puff, bad bad, b*tch *ss, *ss foo, d*ck *ss, ignorant b*tch, hoe hoe, *ss bio, nasty b*tch, big d*ck
DQE	love, lol, cute, omg, beautiful, haha, good, nice, amazing, pretty, happy, wow, awesome, great, cool, perfect, best, guys, day, time, hahaha, gorgeous, god, pic, girl, people, birthday, tttt, life, man, follow, hair, lmao, hot, yeah, going, happy birthday, wait, better, hope, picture, baby, hey, sexy, ya, damn, sh*t, work, adorable, f*ck
CO	hermoso, sdv, sigo, troco, meu deus, troco likes, lindaaa, eu quero, fofo, perfect body, kinds, music video, girls love, allow, lls, spray, shoulders, wait guys, jet, niners, good sh*t, wie, damnnn, garden, post comments, stalk, rail, captain, belieber, sweety, convo, orders, smash, hahaha true, good girl, spider, au, best night, emotional, afternoon, gallery, degrees, hahahahahahah, oui, big time, por favor, beautiful photo, artwork, sb, drooling

TABLE III: Color-coded bullying bigrams detected in Twitter data by PVC and baselines

Method	Detected Bullying Words Color-Coded by Annotation: Bullying, Likely Bullying, Uncertain, Not Bullying.
PVC	singlemost biggest, singlemost, delusional prick, existent *ss, biggest jerk, karma bites, hope karma, jerk milly, rock freestyle, jay jerk, worldpremiere, existent, milly rock, milly, freestyle, *ss b*tch, d*ck *ss, *ss hoe, b*tch *ss, adore black, c*mming f*ck, tgurl, tgurl sl*t, black males, rt super, super annoying, sl*t love, bap babyz, love rt, f*ck follow, babyz, jerk *ss, love s*ck, hoe *ss, c*nt *ss, *ss c*nt, stupid *ss, bap, karma, *ss *ss, f*ggot *ss, weak *ss, bad *ss, nasty *ss, lick *ss, d*ck s*cker, wh*re *ss, ugly *ss, s*ck *ss, f*ck *ss,
DQE	don, lol, good, amp, f*ck, love, sh*t, ll, time, people, yeah, ve, man, going, f*cking, head, didn, day, better, free, ya, face, great, hey, best, follow, haha, big, happy, gt, hope, check, gonna, thing, nice, feel, god, work, game, doesn, thought, lmao, life, c*ck, help, lt, play, hate, real, today,
CO	drink sh*tfaced, juuust, sh*tfaced tm4l, tm4l, tm4l br, br directed, subscribe, follow check, music video, check youtube, checkout, generate, comment subscribe, rt checkout, ada, follback, marketing, featured, unlimited, pls favorite, video rob, beats amp, untagged, instrumentals, spying, download free, free beats, absolutely free, amp free, free untagged, submit music, untagged beats, free instrumentals, unlimited cs, creative gt, free exposure, followers likes, music chance, soundcloud followers, spying tool, chakras, whatsapp spying, gaming channel, telepaths, telepaths people, youtube gaming, dir, nightclub, link amp, mana

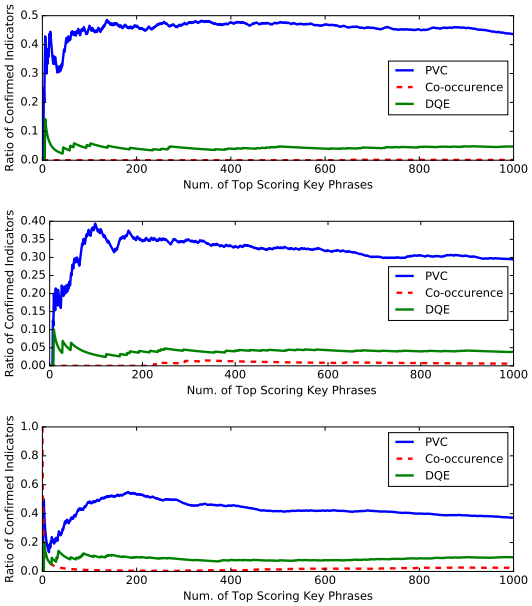

Fig. 2: Precision@k for bullying phrases on Ask.fm (top), Instagram (middle), and Twitter (bottom).

them based on the annotations. Table I, Table II, and Table III list the first 50 words (censored) for Ask.fm, Instagram, and Twitter. The words are color-coded. Using a scoring system of +1 for when an annotator believes the word is a bullying indicator, 0 when an annotator is uncertain, and -1 when an annotator believes the word is not a bullying indicator, we print a word in red if it scored 2 or greater, orange if it scored a 1, gray if it scored a 0, and blue if it scored any negative value. These newly detected indicators suggest that PVC is capable of detecting new offensive words and slang (shown in red).

We inspected the interactions PVC identified in the three datasets and found three categories of note. (1) The first type were represented by bullying conversations containing negative words identified by PVC, but not by other baselines. Two of such cases are shown in Fig. 3. (2) Some cases of conversations contained little prototypical bullying language, such as the slurs from the seed query and beyond. We hypothesize that PVC discovered these because of a combination of discovering new language and considering the typical roles of the conversation participants. In Fig. 4 we show two of these cases. (3) There were interactions that PVC mistakenly identified as harassment, where both we and the annotators consider the interactions be non-harassment. We grouped these false positives into four classes. First, one where users are talking about other people, not addressing each other in their messages. The false positives of this type reveal the importance of considering some keywords like "you," "you are," "your," etc. in our model. One example of such a case is illustrated at the top of Fig. 5. Second, some false positives occur when two users are joking with each other using offensive words, which is common among teenagers, for example. The second conversation in Fig. 5 is one such example. Third, false positives occur when two users have conversation about some negative topics, as shown in the third conversation

User1: lmao don't call me a b*tch. I don't know you, the tweet was just funny, b*tch."
User2: then you @ her and not me you little instigating *ss irrelevant hoe. Run along, b*tch.
User1: hy you mad? Lmao you're irrelevant as f*ck, b*tch. You can get out of my mentions you're a piece of sh*t.
User2: When dumb random *ss irrelevant hoes mention me, they get response. Now get your c*nt *ss on somewhere bruh '

User1: IS A FAKE *SS B*TCH WHO DOESNT DESERVE A MAN'
User2: b*tch you gave a BJ to a manager at McDonalds so you could get a free BigMac'
User1: B*TCH YOU SRE CONFUSING ME WITH YOUR MOTHER'
User2: YOUR MOM HAS BEEN IN THE PORN INDUSTRY LONGER THAN I HAVE BEEN ALIVE'
User1: B*TCH TAKE THAT BACK'
User2: TAKE WHAT BACK?
User1: YOUR RUDE DISRESPECTFUL COMMENTS'
User2: I ONLY SEE THE TRUTH YOU HOE'
User1: TBH DONT GET ME STARTED. I CAN BREAK YOU IN TWO MINUTES.'
User2: DO IT B*TCH, YOUR FAT *SS WILL GET TIRED OF TYPING IN 1 MINUTE'

Fig. 3: Two examples of true positives by PVC that other baselines were not be able to detect. These examples are clearly intense and toxic, but their concentration of obvious swear words may not be high enough for the baseline approaches to identify.

User1: You don't get to call me stupid for missing my point."
User2: I said you're being stupid, because you're being stupid. Who are you to say who gets to mourn whom? Read the link.
User1: You miss my point, again, and I'm the stupid one? Look inwards, f*ckwad.

User1: Stupid she doesnt control the show she cant put it back on you idiot
User1: She isnt going to answer you stupid
User1: Its spelled Carly stupid
User1: She wont answer you stupid

Fig. 4: Examples of harassment detected by PVC and verified by annotators. These examples do not have very obvious offensive-language usage, so methods beyond simple query-matching may be necessary to find them.

in Fig. 5. In this example, the users are discussing sexual promiscuity, and while they are not necessarily being civil to each other, the conversation is not necessarily an example of harassment. Finally, a fourth common form of false positive occurs when no negative words are used in the conversation, but because PVC learned new words it believes to be offensive, it flags these conversations. One example is shown at the bottom of Fig. 5, where there is nothing particularly obvious about the conversation that should indicate harassment.

We analyze the sensitivity of PVC toward some often targeted groups such as those defined by race, gender, sexual orientation, and religion. Because of bias in social media across these groups, PVC will identify some messages containing

User1: LOL he's so nasty, he has a banana shaped faced
User2: Lmao . No comment
User1: L**** is a whore .
User2: Yeah, I'm over him so he can go whore around on his new girlfriend

User1: your f*cking awesome then (:
User2: damn f*cking right < 333333
User1: pretty good looking and seem cool (:
User2: i seem f*cking awesome

User1: if they act like hoes then they getting called hoes'
User2: that's awful thing to say
User1: what! it's true so if a girls a hoe, acts like a hoe and cheats like a hoe she isn't a hoe?
User2: and what exactly is a hoe? And what do you call men who cheat? Hoes too?'
User1: lets just end this conversation because I feel like you're gonna block me soon and I'd rather not lose another friend
User2: no, I mean, if you can say "if she act like a hoe then she gets called a hoe" then I would like to know what a hoe is'
User1: could mean wh*re, could imply she sleeps around, could mean she's just a evil f*ck face that flirts with you and then goes

User1: are u a vegetarian
User2: my parents are rude and wont let me but i dont like meat rip
User1: same dont like it that much but cant live without chicken :/
User2: i hate chicken what
User1: chicken is lyf wyd :/
User2: the only good thing about chicken are nuggets :/
User1: im not demanding i love all shapes and sizes : /
User2: chicken is gross :/

Fig. 5: Examples of false positives by PVC: interactions indentified by PVC that annotators considered non-harassment and appear correctly labeled. These examples include usage of offensive language but may require sophisticated natural language processing to differentiate from harassing usage.

keywords describing sensitive groups as bullying. These can be problematic because these words may often be used in innocuous contexts. We call these mistakes *unfair false positives*, meaning that non-bullying conversations containing sensitive keywords are falsely identified as bullying. Two of such cases are shown in Fig. 6, where these messages containing the keyword "black" may have been flagged because of their including the word. There might be two reasons why we observe these *unfair false positives*: i) sensitive key phrases describing target groups are included in the seed set, or ii) in the dataset, sensitive key phrases co-occur with seed words. We could address the first case by carefully choosing the seed set such that no sensitive key phrases are included; because otherwise we train the model to treat sensitive keywords as indicators, increasing the rate of unfair false positives. To address the second case, we should change our model by considering fairness in our objective function, which is outside of the scope of this paper; but this idea is part of our future work plan.

User1: Owwww sexy
User1: Lets do that
User1: Black and yellow hard

User1: Beef Or Chicken ? Coke Or Pepsi ? White Or Black ? Mercedes Or BMW ? Friendship Or Love ? Yummy! Or Tasty ? Traditional Or Love Marriage ? Sister Or Brother ? Action Or Comedy ? Sweet Or Sour ? Chocolate Or Vanilla ? Strawberry Or Raspberry ? Lemon Or Orange ? Money Or Health?
User2: chicken', ' coke', ' grey;)', ' Mercedes', ' love', ' tasty', ' traditional', ' neither', ' comedy', ' sour', ' chocolate', ' raspberry', ' lemon', ' both probably:)

Fig. 6: Two examples of non-bullying conversations mistakenly flagged by PVC containing the keyword "black".

E. Bully and Victim Score Analysis

While our proposed model learns parameters that represent the tendencies of users to bully or to be victimized, it does not explicitly model the relationship between these tendencies. We can use the learned parameters to analyze this relationship. We plot users based on their bully and victim scores to observe the distributions of the bully and victim scores. We standardize the scores to be between 0 and 1, and in Fig. 8, we show the scatter plot of Twitter users according to their learned bully and victim scores as well as the heatmap plot (two-dimensional histogram) to see how dense the populations are in different regions of bully-victim space. The redder the region, the more users have bully and victim scores in the region. In the heatmap, we can see four hotspots: (1) pure bullies, seen as the horizontal cloud, (2) pure victims, seen as the vertical cloud, (3) victimized bullies, seen as the diagonal cloud, and finally, (4) the more dense hotspot is the region with low bully and victim scores. The existence of these hotspots suggests that most of the users in our Twitter data are not involved in bullying, but those that do have a fairly even mix of being bullies, victims, and both. The heatmap plot for Instagram and Ask.fm are also shown in Fig. 9. In Fig. 7 we show a sample of conversations involving a user with a high bully score (top) and a user with high victim score (bottom). The bully is sending toxic messages to multiple users, and they receive negative response messages as well. The victim, on the other hand, is targeted by three different apparent bullies.

There are also some examples of false positive cases where users are learned to have high bully and victim scores. In one case from Ask.fm shown in Fig. 10, a user receives many messages or a few long messages with many offensive words, but the message is not bullying. Instead, users appear to be using strong language in a positive manner.

F. Bully and Victim Network Analysis

We computed the average in-degree and out-degree of the top 3,000 bullies and victims as shown in Table IV. According to the statistics: 1) the in-degree of these top bullies is less than the in-degree of top victims; 2) the in-degree of top bullies is less or equal than their out-degree; and 3) the in-degree of top victims is greater than or equal to their out-degree. These trends suggest that, on average, high-scoring victims receive

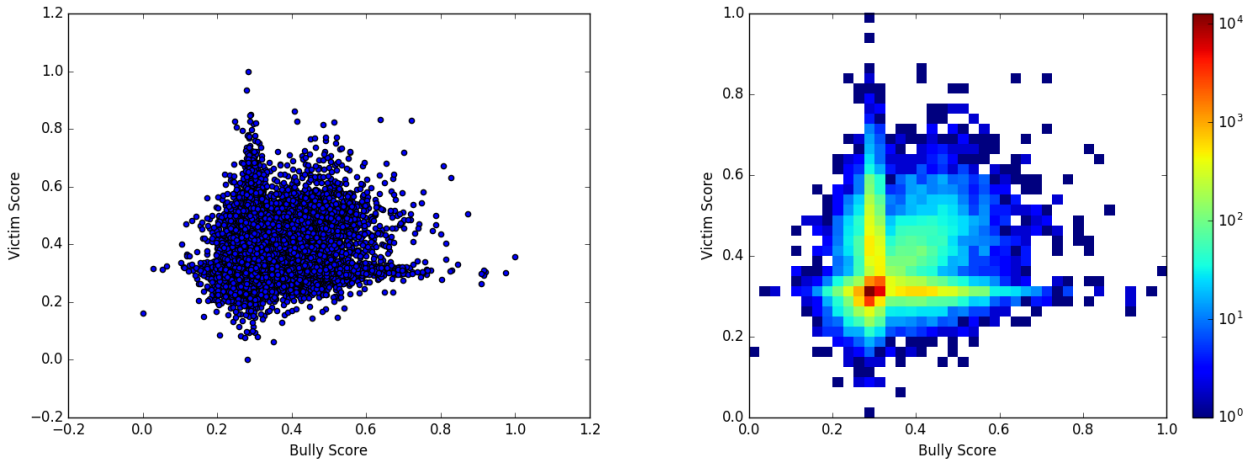

Fig. 8: Scatter (left) and heatmap (right) plots for Twitter. The plots show the distributions of the bully and victim scores. According to the heatmap, most of the users in our Twitter data are not involved in bullying because most of the users occupy the redder region. However, there are moderate number of bullies, victims, and victimized bullies.

more messages than high-scoring bullies; they also receive more messages than they send. In contrast, bullies are more senders of messages than receivers.

We also compute the number of top bully and top victim neighbors for each top bully and top victim. In our Twitter data, around 54.63% of top victims have one bully neighbor, while less than 1% of them have two bully neighbors. Around 4% of bullies have one victim neighbor, and less than 0.2% of them have two or three victim neighbors. On Ask.fm, around 4% of victims have one bully neighbor, while 0.5% of them have two or three bully neighbors. Around 2.5% of bullies are surrounded by one victim, while less than 0.8% of them are surrounded by two and three victims. On Instagram, however, the scale is much smaller. Only 14 victims (out of 3,000) are neighboring by at least one top bully. Less than 4% of bullies have one victim neighbor, and less than 1% of them have between two to five victim neighbors. In general, in most detected bullying cases, there is only one bully and one victim in the immediate social graph.

Average Score	Twitter	Ask.fm	Instagram
Average in-degree of bullies	1.081	6.578	0.011
Average out-degree of bullies	2.286	6.578	2.154
Average in-degree of victims	2.181	7.385	100.99
Average out-degree of victims	1.329	7.385	10.29

TABLE IV: The average in-degree and out-degree of top-scoring bullies and victims. On average, bullies are sending more messages than victims. They also send more messages than they receive, unlike victims who are more receivers of the messages than senders.

To gain a better intuition about the network structure among bullies and victims, we illustrate the communication graph among some detected bullies and victims. To extract the bullying subgraphs, we use a depth-2 snowball sampling starting from detected bully and victim users. Since each node might have hundreds of neighbors, we randomly select at most 10 neighbors and collect the subgraph of the second-hop neighbors subject to this random downsampling. Bullies, victims, and bystanders are shown in *red*, *green*, and *gray*, respectively. We illustrate different communities of such users in Instagram in Figure 11. In the figure, victims are surrounded by several bullies as well as bystanders. This pattern aligns with the idea that a victim could be targeted by a group of bullies. We also observe that, in most cases, bullies are indirectly connected to each other through bystanders or victims. Another interesting point is that not all of a bully's neighbors are victims. In other words, a bully could interact with different people, but they may not bully all of them.

In Figure 12, we show two sub-graphs from Ask.fm. In the top network, there are two bullies and three victims. The bully at the center has three victim neighbors, and one bully neighbor, showing they might have harsh conversations containing indicator key phrases. In the bottom network, the bully user has two victim neighbors. One victim is interacting with their neighbors, while the other one only communicates with the bully user. In general, there are varieties of structures among users: Victims who are bullied by multiple users, bullies who targeting some of their neighbors but not all of them; bullies and victims with many neighbors; or bullies and victims with one neighbor. Examples of these patterns exist within the subgraphs in Figs. 11 and 12.

G. Fairness Analysis

Social media data carries social bias against various demographic groups. Machine learning algorithms trained on this data, therefore, perpetuates this discrimination causing unfair decision-making. Our algorithms are trained on social

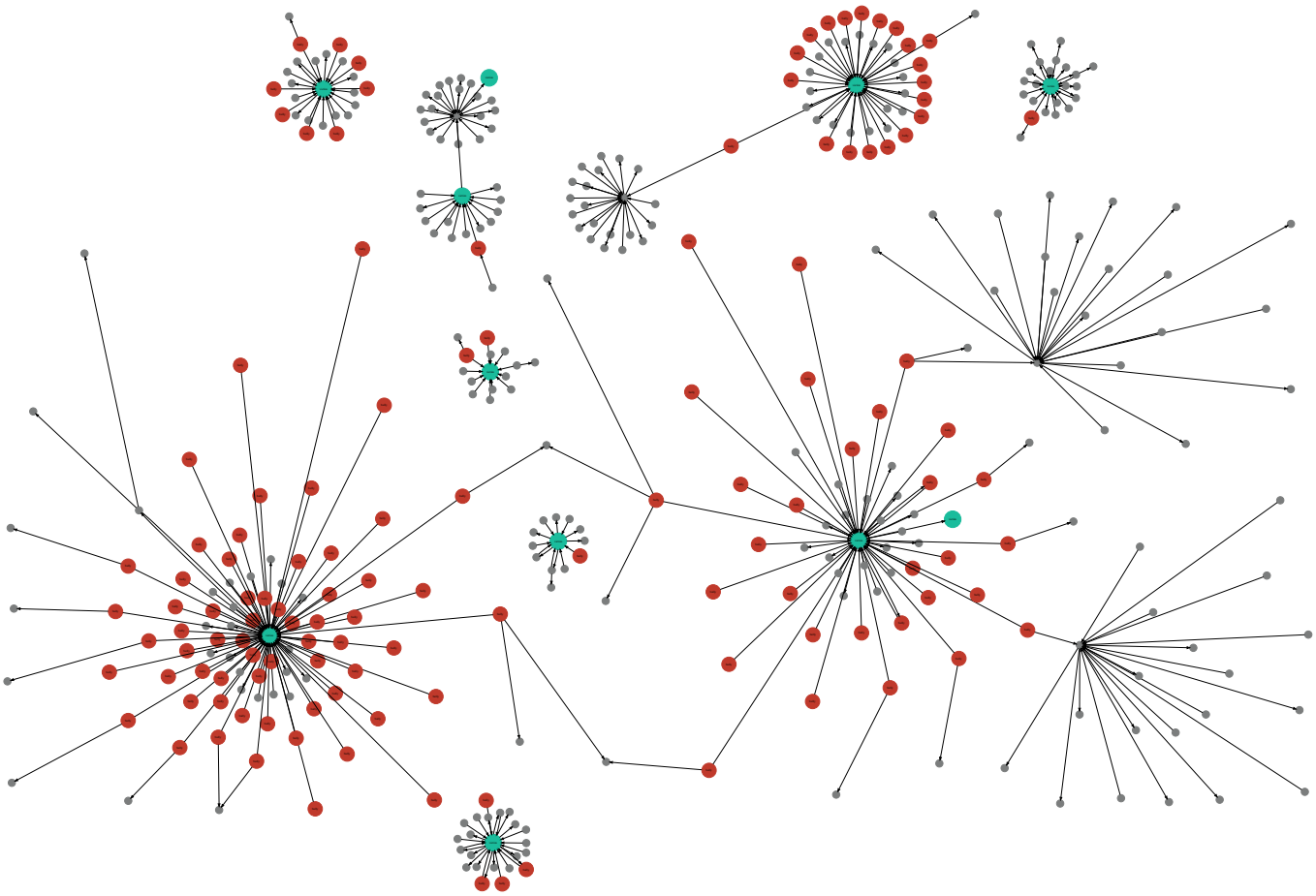

Fig. 11: A sub-graph of bullies and victims in Instagram. *Red*, *green*, and *gray* represent bullies, victims, and bystanders, respectively. Victims have several bully and bystander neighbors. In addition, a bully may interact with different people, but they may not bully all of them.

data; therefore, they might encode society's bias across various demographic groups.

In Fig. 13, a conversation from Instagram is shown in which a user is harassing the other using sensitive keywords describing race and gender, in addition to offensive words. In this section, we examine how fair PVC is toward particular groups. First, we created a list of sensitive keywords describing social groups of four types: race, gender, sexual orientation, and religion. We removed all sensitive keywords from our seed words. Then, we train the PVC based on fair seed words, then we sort the words according to their learned word scores. We then plot the computed word rank of the sensitive keywords. Figure 14 plots the rank of sensitive keywords for Twitter, Instagram, and Ask.fm. Out of 1,425,111 unigrams and bigrams on Twitter, "boy" has the top rank (870) among the sensitive keywords, while "latina" has the lowest rank (184,094). On Instagram and Ask.fm, the rank of the word "queer" is the highest, while the rank of "sikh" and "heterosexual" are the lowest (On Instagram, out of 3,569,295 words, rank of "queer" is 1,973, while rank of "sikh" is 209,189. On Ask.fm, among 1,960,977 words, ranks of "queer" and "heterosexual" words are 677 and 158,747, respectively). These numbers in the plot indicate that the sensitive keywords are spread, with a few appearing among the most indicative bullying words and others appearing much lower in the ranking. It is worth pointing out that only "boy" on Twitter and "queer" on Ask.fm have listed among the top 1,000 bullying phrases (refer to Fig. 2).

Among the selected sensitive keywords, the highest ranked words are *boy* when PVC is trained on Twitter and *queer* when trained on Instagram and Ask.fm. The second highest ranked in all of the datasets is *gay*. The third highest ranked word for Twitter is *black* and in Instagram and Ask.fm is *lesbian*. Comparing gendered words, the rank of *girl* for Twitter is lower than the rank of *boy*; while for both Instagram and Ask.fm, the rank of *girl* is higher than the rank of *boy*. We hypothesize that this happens because in our Twitter data, the word *boy* almost always co-occurs with many other offensive words. Instagram and Twitter are more biased toward *girl*. This trend is consistent with the relative rankings of *woman* versus *man*. Examining the top ten keywords, running PVC on the Twitter data results in four of them belonging to the gender category, two in each of the sexual orientation and race categories, and

bully to User1: Bernie sanders sucks c*ck f*ggot
bully to User2: ice soccer sucks f*cking c*ck f*ggot
bully to User3: all your merchandise sucks f*cking c*ck f*ggot I will take the money
bully to User4: I'm pretty sure there was one where he sniff his finger after he burrows his finger up his cornhole
bully to User5: stupid *ss race baiting wigger you're not fooling anybody. Get cancer
User6 to bully: All them hood n*ggas busted all over your mom like Moses, letting all their people go wigger.
User2 to bully: chill beaner
User1 to bully: But I'm righ
bully to User1: get mad f*ggot
bully to User5: you're not black stop acting. You are race baiting f*ggot and your mother is a mudshark wh*re who let the whole squad hit it raw
User7 to bully I wish I looked like that still
bully to User7: you look like a f*cking beast
User7 to bully: u look like a f*ggot
User8 to bully: idk it's hard to believe he's actually that much of a loser lol
User8 to bully: lol all the sh*t she talked on smokey was probably her just going off what the trolls said
User8 to bully: well I got that n*ggas info, his moms u tryna have some fun? Lol
User8 to bully: aye u know about
User8 to bully: all the sh*t the trolls said ended up being true smh...

User1 to victim: s*ck a d*ck f*ggot
User2 to victim: If you don't show up in my mentions crying like a huge f*ggot then I won't call you the f*ggot word. F*ggot
User2 to victim: Why don't you cry about it, you huge f*ggot. Cry because somebody disagrees with you.
User3 to victim: African diseases would sicken you even more.
victim: to User2 WOOOO man you morons out are coming out of the woodworks today. You sicken me.

Fig. 7: Conversation of users with high bully score (top) and high victim score (bottom). The user with a high bully score is sending toxic messages to other users; the user with a high victim score is receiving intense negative messages while responding to some messages reacting the bullies.

one in the religion category. Using Instagram data, six out of the ten top keywords describe sexual orientation, three describe gender, and one describes religion. Using Ask.fm, five of the highest ranked keywords are about sexual orientation, three are about race, and two are about gender. Overall, these results may indicate that our Twitter data is more biased about gender, while Instagram and Ask.fm are more biased about sexual orientation. The fact that *queer* appears among the highest ranked sensitive keywords may be a result of its history of being used as a slur that has been reclaimed by the LGBT community. While it is now generally accepted to be simply a descriptive word for a group of people, it is also still often used as a slur. Overall, these analyses provide some assurance that the learned PVC models are not overly reliant on these sensitive keywords, but more study is necessary, and we are planning future work with explicit fairness-encouraging learning objectives.

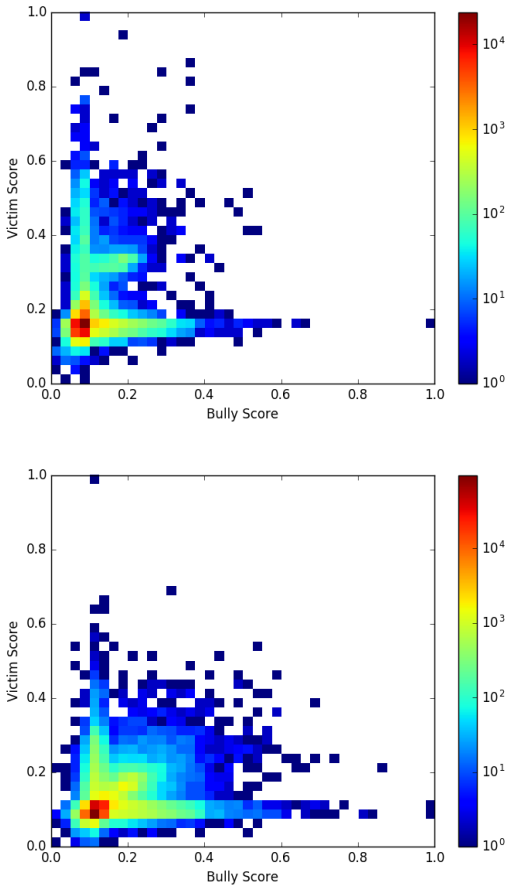

Fig. 9: Heatmap plots for Instagram (top) and Ask.fm (bottom). In both datasets, users with low bully and low victim scores fall in the denser region (red color), but there are users with high bully and high victim scores to a lesser extent.

V. Conclusion

In this paper, we proposed a weakly supervised method for detecting cyberbullying. Starting with a seed set of offensive vocabulary, participant-vocabulary consistency (PVC) simultaneously discovers which users are instigators and victims of bullying, and additional vocabulary that suggests bullying. These quantities are learned by optimizing an objective function that penalizes inconsistency of language-based and network-based estimates of how bullying-like each social interaction is across the social communication network. We ran experiments on data from online services that rank among the most frequent venues for cyberbullying, demonstrating that PVC can discover instances of bullying and new bullying language. In our quantitative analysis, we compute the precision using post-hoc human annotation to evaluate the detected conversations and key phrases by PVC. In our qualitative analysis, we examined discovered conversations, and classified them into some categories of note. Furthermore, we showed some statistics about bullies and victims as well as the distributions of bully and victim scores. We also showed the network structure between some bullies and victims to visualize the social relation

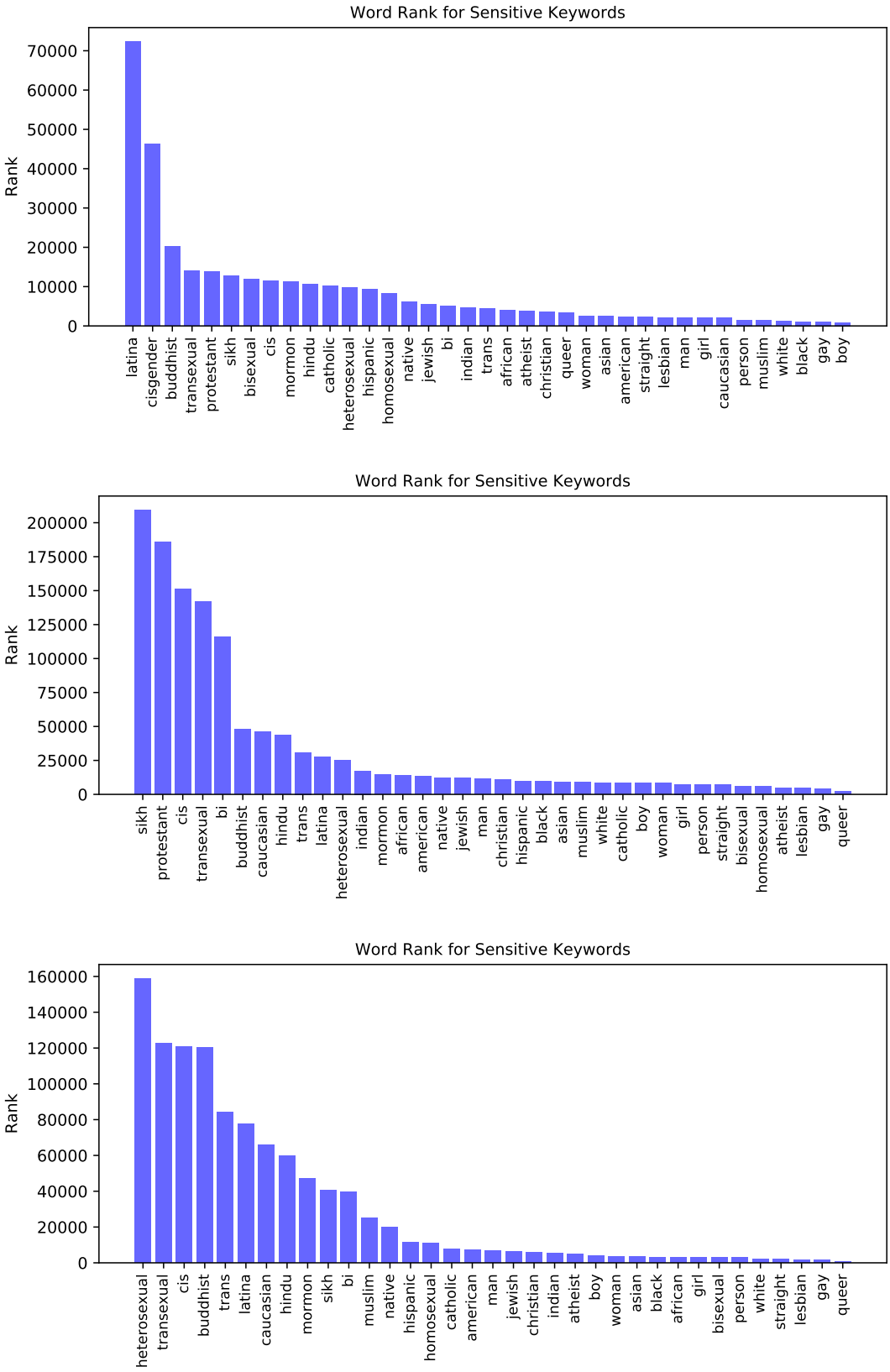

Fig. 14: Sensitive keywords ranks obtained by PVC for Twitter (top), Instagram (middle), and Ask.fm (bottom). The highest word rank in Twitter, Instagram, and Ask.fm is 870 for the word “boy”, 1,973, and 677, respectively for the word “queer”.

User1: I love you so much. I really do. I'm. Just. So. Lesbian. For. You. And were going to go ham when your outta summer school babe :*.
User2: Okay personal. I think you and I would be a boom ass couple.. But ya know
User1: I thought we were going to take it slow between us baby?
User2: Thoughts love?:)
User1: I f*cking love you

Fig. 10: An example of a user with high victim score and low bully score. The user receives messages containing offensive words, but they are not used in an aggressive or harassing manner, and the user does not appear to be a victim.

between the two. Motivating our future work of developing methods to train fair language-based detectors, we tested the sensitivity and bias of PVC toward particular social groups.

Our contribution aims to improve automated detection of cyberbullying. Many more ideas can be incorporated to improve the ability of computers to perform detection. E.g., we are currently developing weakly supervised approaches that additionally consider network features or the sequence of conversations. However, automated detection is only one problem among many that must be solved to adequately address the cyberbullying phenomenon. What to do when cyberbullying is detected is an important open problem. Providing detected bullies and victims advice, filtering content, or initiating human intervention are possible actions an automated system could take upon detection, but how to do any of these tasks in a manner that truly helps is a key open question.

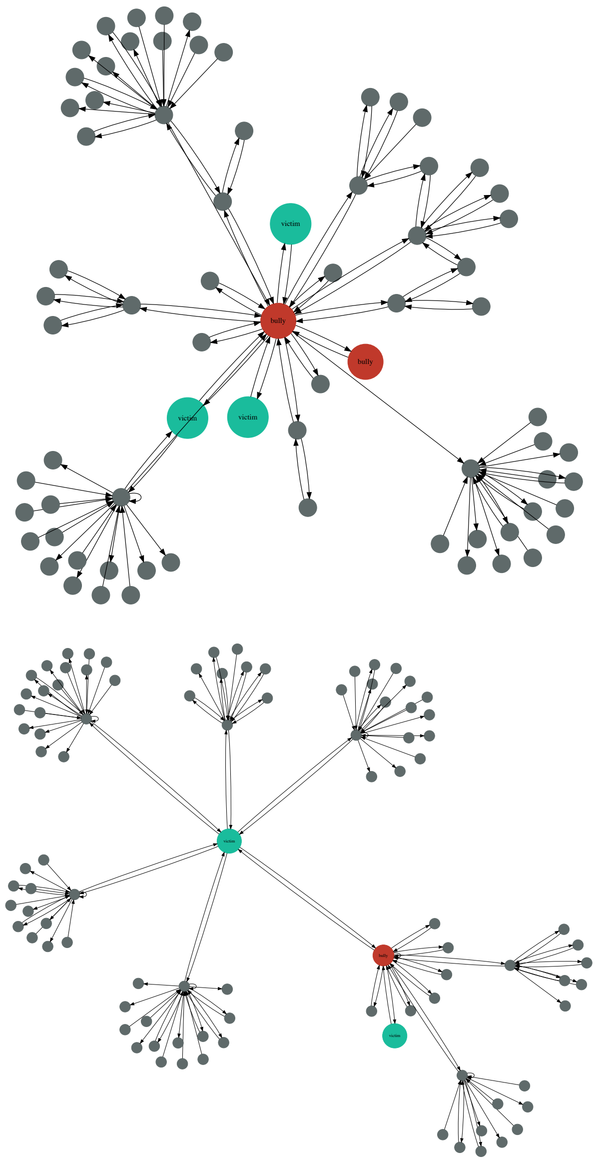

Fig. 12: Two sub-graphs of bullies and victims in Ask.fm. *Red*, *green*, and *gray* represent bullies, victims, and bystanders, respectively. Top: a bully at the center has bully, victim, and neutral neighbors. Bottom: a bully has two victim neighbors; one of them has a large community of friends, while the other is isolated.

References

[1] J. Wang, R. J. Iannotti, and T. R. Nansel, "School bullying among US adolescents: Physical, verbal, relational and cyber," *Journal of Adolescent Health*, vol. 45, pp. 368–375, 2009.

[2] S. C. Herring, "Cyber violence: Recognizing and resisting abuse in online environments," *Asian Women*, vol. 14, pp. 187–212, 2002.

[3] d. boyd, *It's Complicated*. Yale University Press, 2014.

[4] J. S. Donath, "Identity and deception in the virtual community," *Communities in Cyberspace*, vol. 1996, pp. 29–59, 1999.

[5] P. Shachaf and N. Ha, "Beyond vandalism: Wikipedia trolls." *Journal of Information Science*, vol. 36, pp. 357–370, 2010.

[6] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," *CoRR*, vol. abs/1504.00680, 2015.

[7] 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*, vol. 5, no. 2, pp. 245–255, 2015.

[8] I. Whitney and P. K. Smith, "A survey of the nature and extent of bullying in junior/middle and secondary schools," *Educational Research*, vol. 35, no. 1, pp. 3–25, 1993.

[9] D. P. Farrington, "Understanding and preventing bullying," *Crime and Justice*, pp. 381–458, 1993.

[10] J. W. Patchin and S. Hinduja, *Cyberbullying Prevention and Response: Expert Perspectives*. Routledge, 2012.

[11] R. M. Kowalski, S. P. Limber, and P. W. Agatston, *Cyberbullying: Bullying in the Digital Age*. John Wiley & Sons, 2012.

[12] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: Its nature and impact in secondary school pupils," *Journal*

User1: U gay
User1: f*ck you spic
User1: Fagget ass typical spic bitch

User1: U gay
User1: - - is the gayest peoples I have met In my life
User1: - - - u r a f*ggot u can go suck ur dads cock u littles p*ssy f*ggot ur probably on 9 years old bitch IAM 14 dumb hoe f*ggot black bitch suck my cock itll go down ur throat and out of ur mouth u f*ggot black p*ssy
User1: YO BLACK BITCH SHUT YOUR LITTLE BLACK MOUTH ur black cousin can suck my cock too that little bitch probably couldnt fight for sh*t u little black MOTHER F*CKER WHY DONT U GO F*CK UR COUSIN U LITTLES BLACK P*SSYLET U CAN SUCK UR COUSINS DICK TOO BUT THAT SHIT WONT FIT IN YOUR BLACK LITTLE MOUTH I WILL F*CKING HACK UR LITTLE BLACK ASS AN MAKE U SUCK UR DADS DICK SO I SUGGEST U SHUT THE F*CK UP U LITTLE BLACK P*SSY FACE COCK SUCKIN BLACK BITCH SUCK MY DICK AND HAVE A NICE DAY and yo - - - u r unpredictibably retarded and black and suck ur dads an cousins cock u little black bitch
User1: gymnastics 18 lol
User1: - - - shut the f*ck up I will f*cking slap and beat the shit out of u dumbass black little hoe why dont u go f*ck ur cousin he will f*ck ur black butt crack u littles f*ggot and sorry to all the other black people out there these two r just really big d*ck faces

Fig. 13: Two examples of bias toward race and sexual orientation in Instagram. In top example, bully is harassing victim using racist slur ("spic"). In the bottom example, the user is bullying the other one using negative words as well as sensitive keywords about race and sexual orientation ("black" and "gay").

of Child Psychology and Psychiatry, vol. 49, no. 4, pp. 376–385, 2008.

[13] R. S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization," *Computers in Human Behavior*, vol. 26, no. 3, pp. 277–287, 2010.

[14] N. Dordolo, "The role of power imbalance in cyberbullying," *Inkblot: The Undergraduate J. of Psychology*, vol. 3, 2014.

[15] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC," *Mahway: Lawrence Erlbaum Associates*, 2001.

[16] ditchthelabel.org, "The annual cyberbullying survey," *http://www.ditchthelabel.org/*, 2013.

[17] M. Dadvar, F. de Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," *Dutch-Belgian Information Retrieval Workshop*, pp. 23–25, February 2012.

[18] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," *Intl. Conf. on Social Computing*, pp. 71–80, 2012.

[19] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," *ICWSM Workshop on Social Mobile Web*, 2011.

[20] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," *International Conference on Machine Learning and Applications and Workshops (ICMLA)*, vol. 2, pp. 241–244, 2011.

[21] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on Web 2.0," *Content Analysis in the WEB 2.0*, 2009.

[22] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," *Communications in Information Science and Management Engineering*, vol. 3, no. 5, pp. 238–247, May 2013.

[23] M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, and K. Araki, "Machine learning and affect analysis against cyber-bullying," in *Linguistic and Cognitive Approaches to Dialog Agents Symposium*, 2010, pp. 7–16.

[24] H. Margono, X. Yi, and G. K. Raikundalia, "Mining Indonesian cyber bullying patterns in social networks," *Proc. of the Australasian Computer Science Conference*, vol. 147, January 2014.

[25] Q. Huang and V. K. Singh, "Cyber bullying detection using social and textual analysis," *Proceedings of the International Workshop on Socially-Aware Multimedia*, pp. 3–6, 2014.

[26] N. Tahmasbi and E. Rastegari, "A socio-contextual approach in automated detection of cyberbullying," in *Proceedings of the 51stHawaii International Conference on System Sciences*, 2018, pp. 2151–2160.

[27] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Detecting aggressors and bullies on Twitter," in *Proceedings of the 26th International Conference on World Wide Web Companion*, ser. WWW '17 Companion, 2017, pp. 767–768. [Online]. Available: https://doi.org/10.1145/3041021.3054211

[28] V. K. Singh, Q. Huang, and P. K. Atrey, "Cyberbullying detection using probabilistic socio-textual information fusion," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, vol. 00, Aug. 2016, pp. 884–887. [Online]. Available: doi.ieeecomputersociety.org/10.1109/ASONAM.2016.7752342

[29] A. Bellmore, A. J. Calvin, J.-M. Xu, and X. Zhu, "The five W's of bullying on Twitter: Who, what, why, where, and when," *Computers in Human Behavior*, vol. 44, pp. 305–314, 2015.

[30] Z. Ashktorab and J. Vitak, "Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers," in *Proc. of the CHI Conf. on Human Factors in Computing Systems*, 2016, pp. 3895–3905.

[31] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on twitter," *CoRR*, vol. abs/1702.06877, 2017. [Online]. Available: http://arxiv.org/abs/1702.06877

[32] ——, "Measuring #gamergate: A tale of hate, sexism, and bullying," *CoRR*, vol. abs/1702.07784, 2017. [Online]. Available: http://arxiv.org/abs/1702.07784

[33] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and American Academy of Child & Adolescent Psychiatry, "Hate is not binary: Studying abusive behavior of #gamergate on twitter," *CoRR*, vol. abs/1705.03345, 2017. [Online]. Available: http://arxiv.org/abs/1705.03345

[34] C. Chelmis, D. Zois, and M. Yao, "Mining patterns of cyberbullying on twitter," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, vol. 00, Nov. 2018, pp. 126–133. [Online]. Available: doi.ieeecomputersociety.org/10.1109/ICDMW.2017.22

[35] D.-S. Zois, A. Kapodistria, M. Yao, and C. Chelmis, "Optimal online cyberbullying detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE SigPort, 2018. [Online]. Available: http://sigport.org/2499

[36] T. Davidson, D. Warmsley, M. W. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *CoRR*, vol. abs/1703.04009, 2017. [Online]. Available: http://arxiv.org/abs/1703.04009

[37] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra, "Towards understanding cyberbullying behavior in a semi-anonymous social network," *IEEE/ACM International Conf. on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 244–252, August 2014.

[38] H. Hosseinmardi, S. Li, Z. Yang, Q. Lv, R. I. Rafiq, R. Han, and S. Mishra, "A comparison of common users across Instagram and Ask.fm to better understand cyberbullying," *IEEE Intl. Conf. on Big Data and Cloud Computing*, 2014.

[39] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of cyberbullying incidents on the Instagram social network," *Association for the Advancement of Artificial Intelligence*, 2015.

[40] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Workshop on Language in Social Media*, 2012, pp. 19–26.

[41] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *International Conference on World Wide Web*, 2015, pp. 29–30.

[42] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings Intl. Conf. on World Wide Web*, 2016, pp. 145–153.

[43] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski, "Learning to identify internet sexual predation," *Intl. J. of Electronic Commerce*, vol. 15, no. 3, pp. 103–122, 2011.

[44] D. U. Patton, K. McKeown, O. Rambow, and J. Macbeth, "Using natural language processing and qualitative analysis to intervene in gang violence," *arXiv preprint arXiv:1609.08779*, 2016.

[45] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.

[46] K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp, "Incorporating query expansion and quality indicators in searching microblog posts," *Proc. of the European Conference on Advances in Information Retrieval*, vol. 15, no. 5, pp. 362–367, November 2011.

[47] V. Lavrenko and W. B. Croft, "Relevance based language models," in *Proc. of the International ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2001, pp. 120–127.

[48] A. Mahendiran, W. Wang, J. Arredondo, B. Huang, L. Getoor, D. Mares, and N. Ramakrishnan, "Discovering evolving political vocabulary in social media," in *Intl. Conf. on Behavioral, Economic, and Socio-Cultural Computing*, 2014.

[49] E. Raisi and B. Huang, "Cyberbullying identification using participant-vocabulary consistency," *CoRR*, vol. abs/1606.08084, 2016. [Online]. Available: http://arxiv.org/abs/1606.08084

[50] ——, "Cyberbullying detection with weakly supervised machine learning," in *Proceedings of the IEEE/ACM International Conference on Social Networks Analysis and Mining*, 2017.

[51] noswearing.com, "List of swear words & curse words," *http://www.noswearing.com/dictionary*, 2016.

[52] A. Bifet and E. Frank, "Sentiment knowledge discovery in Twitter streaming data," *Intl. Conf. on Discovery Science*, pp. 1–15, 2010.

[53] T. H. Silva, P. O. de Melo, J. M. Almeida, J. Salles, and A. A. Loureiro, "A picture of Instagram is worth more than a thousand words: Workload characterization and application," *DCOSS*, pp. 123–132, 2013.

[54] 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 *Intl. Conf. on Social Informatics*, 2015, pp. 49–66.

[55] N. Ramakrishnan, P. Butler, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz, C. Kuhlman, A. Marathe, L. Zhao, H. Ting, B. Huang, A. Srinivasan, K. Trinh, L. Getoor, G. Katz, A. Doyle, C. Ackermann, I. Zavorin, J. Ford, K. Summers, Y. Fayed, J. Arredondo, D. Gupta, and D. Mares, "'Beating the news' with EMBERS: Forecasting civil unrest using open source indicators," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2014, pp. 1799–1808.