Cyberbullying Detection with Weakly Supervised Machine Learning

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*—Detrimental online behavior such as harassment and cyberbullying is becoming a serious, large-scale problem damaging people's lives. This phenomenon is creating a need for automated, data-driven techniques for analyzing and detecting such behaviors. We propose a 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, the learning algorithm only requires weak supervision. Experts provide a small seed vocabulary 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). We evaluate PVC on three social media data sets, demonstrating quantitatively and qualitatively its effectiveness in cyberbullying detection.**

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, but these approaches require annotators to provide large amounts of labeled examples, each of which requires consideration of social context and changing language.

The algorithm we present here encodes such complexities into an efficiently learnable model. This algorithm 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.

ASONAM '17, July 31 - August 03, 2017, Sydney, Australia

ACM ISBN 978-1-4503-4993-2/17/07?/$15.00
http://dx.doi.org/10.1145/3110025.3110049

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. 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.

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]. Researchers have applied machine learning methods to better understand social-psychological issues surrounding the idea of bullying [26]. 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 [27].

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 [28]. 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 [29]. They also studied the detection of cyberbullying incidents over images on Instagram, focusing on the distinction between cyberbullying and cyber-aggression [30], 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 [31]–[33], online predation [34], and the analysis of gang activity on social media [35], 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 [36]. Query expansion aims to suggest a set of related keywords for user-provided queries. Massoudi et al. [37] use temporal information as well as co-occurrence to score the related terms to expand the query. Lavrenko et al. [38] 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. [39] 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 [40]. This paper represents the extended and complete description, analysis, and results from the study.

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

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

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

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

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.,

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

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

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

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

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

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

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 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 [41]. 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], [42], [43]. 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. [29]. 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. [44], 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 [45]. 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 [38]) 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

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

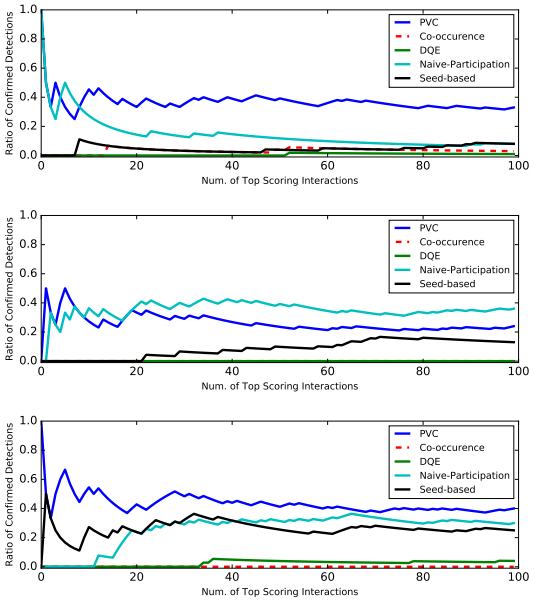

Fig. 1: Precision@k for bullying interactions on Ask.fm (top), Instagram (middle), and Twitter (bottom).

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 Twitter and Ask.fm data, PVC significantly dominates the other methods for all thresholds, while in the Instagram data, it falls below the precision of the naive-participation score. Inspecting some conversations PVC returned for Instagram, we found cases that the annotators did not consider harassment when the conversations and messages were much longer than usual; annotators did not seem to read long messages attentively to recognize the evidence of harassment buried in the middle or the end of the text. We also saw interactions where users are harassed for being a fan of a celebrity. We hypothesize that annotators may have dismissed these conversations when they saw that it was discussing pop culture, not noticing the toxic, hurtful messages within the discussion. 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.

D. Qualitative Analysis

We analyzed the 1,000 highest-scoring, non-seed terms produced by PVC, DQE, and co-occurrence and categorized them based on the annotations. Table I lists the first 50 words (censored) for Ask.fm. The word lists for the other data sets show similar trends and are omitted for space. 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 data sets and found three categories of note. First, we saw some cases of conversations that contained little prototypical bullying language, such as the slurs in the seed query. We hypothesize that PVC discovered these because of a combination of discovering new language and considering the typical roles of the conversation participants. Two of these are shown in Fig. 3. Second, we found cases where PVC seemed to identify evidence of harassment but the annotators disagreed, possibly incorrectly. Two of these are shown in Fig. 4. These cases included conversations that had long messages, which required the annotators to pay more attention to find the often subtle evidence of harassment, or conversations that may have been on the border of the definition of harassment, such as trolling of celebrity accounts. Third, we found interactions that PVC mistakenly identified as harassment, where both we and the annotators consider the interactions be non-harassment. These examples, two of which are shown in Fig. 5, often include typical harassment language, such as the case shown where a user is discussing the offensiveness of such language.

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

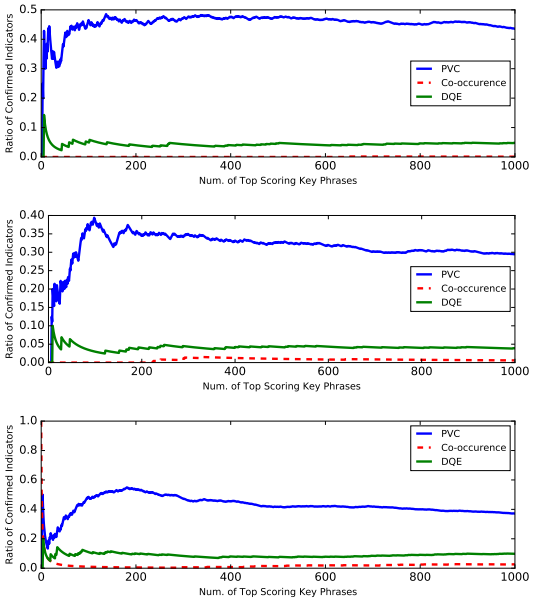

Fig. 2: Precision@k for bullying phrases on Ask.fm (top), Instagram (middle), and Twitter (bottom).

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

User1: Truth is. You hate me. Rate- my mom said if I have nothing nice to say, I shouldn't say anything at all.
User2: Let me explain why I hate you. Okay so I only hate three people so obviously you have pissed me off enough to get on that list. So for starters, you obviously said you think that T*** and J*** will go to hell. Don't say two of best friends will go to hell because who else would T and J be? Second, you called R*** gay. That's not acceptable either. He even had a girlfriend at the time. You blamed it on your friend P**** or whatever her name is. So you didn't accept what you did and tried to hide it and that didn't work because we ALL know you called him gay multiple times. Another thing is, you are honestly so ignorant and arrogant. You think you are the best of the best and think you have the right to do whatever you want, whenever you want but you cant. I hate to break it to you, but you aren't the little princess you think you are. and you are basically calling me ugly in that rate. But you know what? i know im not the prettiest but at least im not the two-faced, conceited, b*tch who thinks that they can go around saying whatever they want. because saying people will go to hell can hurt more than you think. calling someone gay is really hurtful. youve called me ugly plenty of times, too. so congratulations you have made it on the list of people i hate. and i could go on and on but i think ill stop here. btw; your mom obviously didnt teach you that rule well enough. "buh-bye"

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.

Fig. 3: 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 are necessary to find them.

User1: Making fun of us for being gay isn't a joke, it doesn't change the fact that you're still f*cking disgusting.
User2: I'm not calling him a gay because hes gay you f*cking retard, my bestfriend is gay and I have 100
User1: You liked his like for an insult, so obviously he was going to insult you god you're f*cking brain dead.
User2: why dont you shut the f*ck up you little f*ck face. Being Braindead isn't something to joke about you immature little f*cking nuisance. Lol go cry to your little f*ggot boyfriend and get the f*ck off my ask.
User1: leave my boyfriend alone I love him to pieces and hes a beauty. Ive never met someone so nice and so beautiful I honestly dont know why you're hating on him hes a stud and you're actually gross... M*** is the nicest most down to earth boy I've ever met :)
User2: lol, maybe you need to get youre f*cking eyes checked, and see he said sh*t to me first k? mind your own f*cking buisness lol, and again a beauty? HAHAHAHAHAHAHAHA ya and youre the f*cking beast. Man whats with all the old people on here jesus.

User1: guys I think we should apologize to you know why Because shes so pathetic and weak she has to pick on other people that was doing absolutely nothing to her We should feel sorry for her Shes probably so sad and tired of her life she has to pick on other people Pathetic people that picks on random people she doesnt even know is so sad Im being a hypocrite for picking on her I guess but come on by what I read shes a sad little girl whos insecure and lonely and needs to learn how to spell.
User1: sigh Im ugly okay yeah maybe Im not the prettiest person ever but at least my personality isnt sh*t ugly like you tw*t face You little f*ck face should go back to elementary school if YOU think I have bad spelling And seriously if Im ugly it seriously doesnt make you any better Youre a attention wh*re who seriously needs to learn how to spell and stop being a little tw*t

Fig. 4: Examples of possible false negatives by annotators: interactions detected by PVC, labeled by annotators to be non-harassment, but appear to include evidence of harassment hidden within long conversations.

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.

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.

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,

User1: 'why are women still being subjected to abuse, FGM, cat-calling, sl*t-shaming and rape in a far greater
User2: Cat calling isn't that bad and sl*t shaming is when a girl sleeps with a bunch of guys right?
User1: the point is, there is no male equivalent - maybe stud? Nice.
User2: what about these? Man sl*t, pervert, trash, d*uche bag, and man wh*re?
User1: man sl*t and wh*re borrow from the original feminine. Pervert; trash are gender neutral. D*uche bag = vaginal wash

User1: People why do you feel the need to be bullies It wasnt okay in school and it isnt okay now People who hate only do so because they hate on the inside its low self esteem But why should celebrities be the target of YOUR pain and self esteem issues Just because they are in the public eye does not give you the right to judge criticize and be nasty If you feel someone is not very attractive even though you dont know them or their heart which is where real beauty comes from just keep it to yourself What do you gain by putting others down Does it make you feel bigger and greater If so its only short term So get help Beauty is in the eye of the beholder and real beauty comes from the inside This is why you judge because you dont feel beautiful inside Well change it get help and change what you dont like Life will be a lot more peaceful and fulfilling if you feel self worth So get it get help and change your life And please stop judging others It doesnt help you and it only hurts others
User1: Thanks everybody I know my comment was long but hopefully it registers with some people and they can think twice before judging Nobody has the right to judge everybody has the right to an opinion but if it is hurtful then why say it We should all just respect each other for who they are and each others differences We are all Gods creations and we are all beautiful in some way even the ones who only show hate and judgment they do it for a reason and obviously it raises their esteem to put someone down But beauty is in the eye of the beholder Love each other for our differences and stop judging just respect each other We dont have to agree but we should respect each other
User1: thanks
User1: Thats disgusting it wasnt funny and to go so far as to say queer and f*ggot DISGUSTING No one should EVER be called the F word it is EXTREMELY insensitive and totally insulting I am not excusing the fact you were called a stupid b*tch that was also uncalled for and insulting But dont stoop so low as to ever say the F word or dis someone for their sexuality would it be insulting if you were called straight Why use a persons sexual orientation as an insult Not cool And NEVER say f*ggot it is extremely hurtful and vile And no I am not gay I am straight However I do believe in equality And I know how disgusting that word is

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.

"Cyberbullying: Its nature and impact in secondary school pupils," *Journal 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] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

[39] 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.

[40] 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

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

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

[43] 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.

[44] 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.

[45] 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.