Cyberbullying Identification Using Participant-Vocabulary Consistency

Elaheh Raisi   ELAHEH@VT.EDU

Virginia Tech, Blacksburg, VA

Bert Huang   BHUANG@VT.EDU

Virginia Tech, Blacksburg, VA

Abstract

With the rise of social media, people can now form relationships and communities easily regardless of location, race, ethnicity, or gender. However, the power of social media simultaneously enables harmful online behavior such as harassment and bullying. Cyberbullying is a serious social problem, making it an important topic in social network analysis. Machine learning methods can potentially help provide better understanding of this phenomenon, but they must address several key challenges: the rapidly changing vocabulary involved in cyberbullying, the role of social network structure, and the scale of the data. In this study, we propose a model that simultaneously discovers instigators and victims of bullying as well as new bullying vocabulary by starting with a corpus of social interactions and a seed dictionary of bullying indicators. We formulate an objective function based on participant-vocabulary consistency. We evaluate this approach on Twitter and Ask.fm data sets and show that the proposed method can detect new bullying vocabulary as well as victims and bullies.

1. Introduction

Social media has significantly changed the nature of society. Our ability to connect with others has been massively enhanced, removing boundaries created by location, gender, age, and race. However, the benefits of this hyper-connectivity also come with the enhancement of detrimental aspects of social behavior. Cyberbullying is an example of one such behavior that is heavily affecting the younger generations (Boyd, 2014). The Cyberbullying

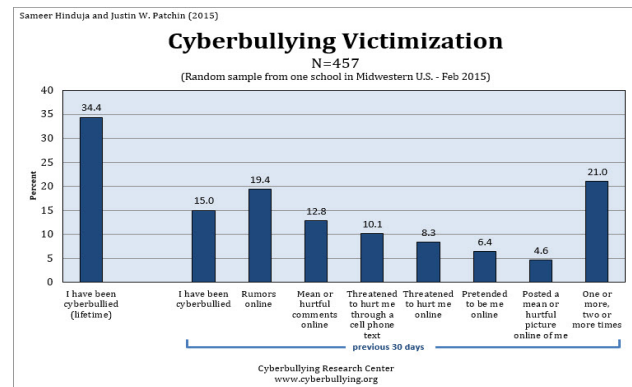*Figure 1.* Survey Statistics on Cyberbullying Experiences. Data collected and visualized by the Cyberbullying Research Center (http://cyberbullying.org/).

Research Center defines cyberbullying as "willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices." Like traditional bullying, cyberbullying occurs in various forms. Examples include name calling, rumor spreading, threats, and sharing of private information or photographs.[1,2] Even seemingly innocuous actions such as supporting offensive comments by "liking" them can be considered bullying (Wang et al., 2009). As stated by the National Crime Prevention Council, around 50% of American young people are victimized by cyberbullying. According to the American Academy of Child and Adolescent Psychiatry, victims of cyberbullying have strong tendencies toward mental and psychiatric disorders (American Academy of Child Adolescent Psychiatry, 2016). In extreme cases, suicides have been linked to cyberbullying (Goldman, 2010; Smith-Spark, 2013). The phenomenon is widespread, as indicated in Fig. 1, which plots survey responses collected from students. These facts make it clear that cyberbullying is a serious health threat.

Machine learning can be useful in addressing the cyberbullying problem. Recently, various studies considered supervised, text-based cyberbullying detection, classifying

---

[1]http://www.endcyberbullying.org
[2]http://www.ncpc.org/cyberbullying

social media posts as 'bullying' or 'non-bullying'. Training data is annotated by experts or crowdsourced workers. Since bullying often involves offensive language, text-based cyberbullying detection studies often use curated swear words as features, augmenting other standard text features. Then, supervised machine learning approaches train classifiers from this annotated data (Yin et al., 2009; Ptaszynski et al., 2010; Dinakar et al., 2011).

We identify three significant challenges for supervised cyberbullying detection. First, annotation is not an easy task. It requires expertise about culture, examination of the social structure of the individuals involved in each interaction. Because of the difficulty of labeling these often subtle distinctions between bullying and non-bullying, there is likely to be disagreement among labelers, making costs add up quickly for a large-scale problem. Second, reasoning about which individuals are involved in bullying should do joint, or collective, classification. E.g., if we believe a message from A to B is a bullying interaction, we should also expect a message from A to C to have an increased likelihood of also being bullying. Third, language is rapidly changing, especially among young populations, making the use of static text indicators prone to becoming outdated. Some curse words have completely faded away or are not as taboo as they once were, while new slang is frequently introduced into the culture. These three challenges suggest that we need a dynamic methodology to collectively detect emerging and evolving slurs with only weak supervision.

In this paper, we introduce an automated, data-driven method for cyberbullying identification. The eventual goal of such work is to detect such harmful behaviors in social media and intervene, either by filtering or by providing advice to those involved. Our proposed learnable model takes advantage of the fact that the data and concepts involve relationships. We train this relational model in a weakly supervised manner, where human experts provide a small seed set of phrases that are highly indicative of bullying. Then the algorithm finds other bullying terms by extrapolating from these expert annotations. In other words, our algorithm detects cyberbullying from key-phrase indicators. We refer to our proposed method as the participant-vocabulary consistency (PVC) model; It 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 are highly indicative bullying phrases in messages between specific users.

A study by ditchthelabel.org (2013) found that Facebook, YouTube, Twitter, and Ask.fm are the platforms that have

the most frequent occurrences of cyberbullying. To evaluate the participant-vocabulary consistency method, we ran our experiments on Twitter and Ask.fm data. From a list of highly indicative of bullying key phrases, we subsample small seed sets to train the algorithm. We then examine the participant-vocabulary consistency method to see how well it recovers the remaining, held-out set of indicative phrases. Additionally, we extract the detected most bullying phrases and qualitatively verify that they are in fact examples of bullying.

2. Related Work

There are two main branches of research related to our topic. One of them is online harassment and cyberbullying detection; the other one is associated with automated vocabulary discovery. Various studies have used fully supervised learning to classify bully posts from non-bully posts. Many of them focus on the textual features of post to identify cyberbullying incidents (Dinakar et al., 2011; Ptaszynski et al., 2010; Hosseinmardi et al., 2015; Chen et al., 2012; Margono et al., 2014). Some of them use other features than only textual features, for example content, sentiment, and contextual features (Yin et al., 2009), the number, density and the value of offensive words (Reynolds et al., 2011), or the number of friends, network structure, and relationship centrality (Huang & Singh, 2014). Nahar et al. (2013) used semantic and weighted features; they also identify predators and victims using a ranking algorithm. Many studies have been applied machine learning techniques to better understand social-psychological issues such as bullying. They used data sets such as Twitter, Instagram and Ask.fm to study negative user behavior (Bellmore et al., 2015; Hosseinmardi et al., 2014a;b).

Various works use query expansion to extend search queries to dynamically include additional terms. For example, Massoudi et al. (2011) use temporal information as well as co-occurrence to score the related terms to expand the query. Mahendiran et al. propose a method based on probabilistic soft logic to grow a vocabulary using multiple indicators (e.g., social network, demographics, and time).

3. Proposed Method

To model the cyberbullying problem, for each user u_i, we assign a bully score b_i and a victim score v_i. The bully score measures how much a user tends to bully others; likewise, victim score indicates how much a user tends to be bullied by other users. For each feature w_k, we associate a feature-indicator score that represents how much the feature is an indicator of a bullying interaction.

Each feature represents the existence of some descriptor in the message, such as n-grams in text data. The sum of senders bullying score and receivers victim score ($b_i + v_i$) specifies the message's *social bullying score*, which our model aims to make consistent with the vocabulary-based feature score. We formulate a regularized objective function that penalizes inconsistency between the social bullying score and each of the feature scores.

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

Learning is then an optimization problem over parameter vectors $\mathbf{b}$, $\mathbf{v}$, and $\mathbf{w}$. The consistency penalties are determined by the structure of the social data. We include information from an expert-provided initial seed of highly indicative bully words. We require these seed features to have a high score, adding the constraint:

$$\min_{\mathbf{b}, \mathbf{v}, \mathbf{w}} J(\mathbf{b}, \mathbf{v}, \mathbf{w}; \lambda) \text{ s.t. } w_k = 1.0, \forall k : x_k \in S. \quad (2)$$

We refer to this model as the participant-vocabulary consistency model because we optimize the consistency of scores computed based on the participants of each social interaction as well as the vocabulary used in each interaction. The objective function Eq. (1) is not jointly convex; However, if we optimize each parameter vector in isolation, we then solve convex optimizations with closed form solutions. The optimal value for each parameter vector given the others can be obtained by solving for their zero-gradient conditions. The update for bully score vector $\mathbf{b}$ is:

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

where $\{m \in M | s(m) = i\}$ 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 closed-form solution for optimizing with respect to victim score vector $\mathbf{v}$ is

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

where $\{m \in M | r(m) = j\}$ is the set of messages sent to user j. The word score vector $\mathbf{w}$ can be updated with

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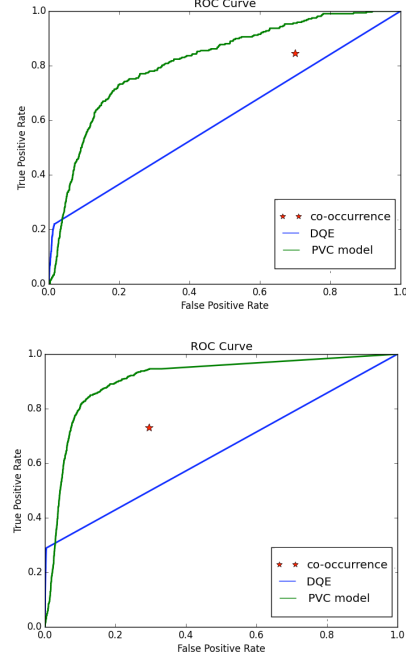

Figure 2. ROC curve of recovered target words for Ask.fm (top) and Twitter (bottom).

The set $\{m \in M | k \in f(m)\}$ indicates the set of messages that contain the kth feature or n-gram. The strategy of fixing some parameters and solving the optimization problem for the rest of parameters known as alternating least-squares. Our algorithm iteratively updates each of parameter vectors $\mathbf{b}$, $\mathbf{v}$, and $\mathbf{w}$ until convergence. The output of the algorithm is the bully and victim score of all the users and the bully score of all the words.

4. Experiments

We ran our experiments on Twitter and Ask.fm data. To collect our Twitter data set, we first collected tweets using words from our curse word dictionary; then we use snowball sampling from these profane tweets to collect additional messages. Since we want to consider messages between users, we are interested in @-replies, in which one user directs a message to another or answers a post from another user. From our full data set, we remove the tweets which are not part of any such conversation, all retweets and duplicate tweets. After the data collection and post-processing, our Twitter data set contains 180,355 users and 296,308 tweets.

Ask.fm is a social question-answering service, where users post questions on the profiles of other users. We use part of the Ask.fm data set collected by Hosseinmardi et al. (2014b). They used snowball sampling, collecting user profiles and a complete list of answered questions. We remove all the question-answer pairs where the identity of the questioner is hidden. Our Ask.fm data set consists of

Table 1. Identified bullying bigrams detected by participant-vocabulary consistency from Twitter and Ask.fm data sets.

Data Set	Selected High-Scoring Words
Twitter	sh*tstain arisew, c*nt lying, w*gger, commi f*ggot, sp*nkbucket lowlife, f*cking nutter, blackowned whitetrash, monster hatchling, f*ggot dumb*ss, *ssface mcb*ober, ignorant *sshat
Ask.fm	total d*ck, blaky, ilysm n*gger, fat sl*t, pathetic waste, loose p*ssy, c*cky b*stard, wifi b*tch, que*n c*nt, stupid hoee, sleep p*ssy, worthless sh*t, ilysm n*gger

41,833 users and 286,767 question-answer pairs.

We compare our method with two baselines. The first is **co-occurrence**. All words or bigrams that occur in the same tweet as any seed word are given the score 1, and all other words have score 0. The second baseline is **dynamic query expansion** (DQE) (Ramakrishnan et al., 2014). DQE extracts messages containing seed words, then computes the *document frequency score* of each word. It then iterates selection of the k highest-scoring keywords and re-extraction of relevant messages until it reaches a stable state.

We evaluate performance by considering held out words from our full curse word dictionary as the relevant target words. We measure the true positive rate and the false positive rate, computing the receiver order characteristic (ROC) curve for each compared method. Fig. 2 contains the ROC curves for both Twitter and Ask.fm. Co-occurrence only indicates whether words co-occur or not, so it forms a single point in the ROC space. Our PVC model and DQE compute real-valued scores for words and generate curves. DQE produces very high precision, but does not recover many of the target words. However, co-occurrence detects a high proportion of the target words, but at the cost of also recovering a large fraction of non-target words. PVC is able to recover a much higher proportion of the target words comparing DQE. PVC enables a good compromise between recall and precision.

We also compute the average score of target words, non-target words, and all of the words. If the algorithm succeeds, the average target-word score should be higher than the overall average. For both Twitter and Ask.fm, our proposed PVC model can capture target words much better than baselines. We measured how many standard deviations the average target-word score is above the overall average. For Twitter, PVC provides a lift of around 1.5 standard deviations over the overall average, while DQE only produces a lift of 0.242. We also observe the same behavior for Ask.fm: PVC learns scores that have a

defined lift between the overall average word score and the average target word score (0.825). DQE produces a small lift (0.0099). Co-occurrence has no apparent lift.

By manually examining the 1,000 highest scoring words, we find many seemingly valid bullying words. These detected curse words include sexual, sexist, racist, and LGBT (lesbian, gay, bisexual, and transgender) slurs. Table 1 lists some of these high-scoring words from our experiments.

The PVC algorithm also computes bully and victim scores for users. By studying the profiles of highly scored victims in Ask.fm, we noticed that some of these users do appear to be bullied. This happens in Twitter as well, in which some detected high scoring users are often using offensive language in their tweets. Fig. 3 shows some bullying comments to an Ask.fm user and her responses, all of which contain offensive language and seem highly inflammatory.
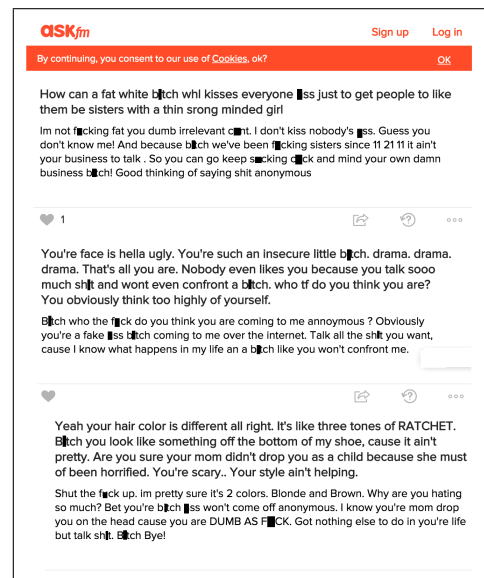

Figure 3. Example of an Ask.fm conversation containing possible bullying and heavy usage of offensive language.

5. Conclusion

In this paper, we proposed the participant-vocabulary consistency method to simultaneously discover victims, instigators, and vocabulary of words indicates bullying. Starting with seed dictionary of high-precision bullying indicators, we optimize an objective function that seeks consistency between the scores of the participants in each interaction and the scores of the language use. For evaluation, we perform our experiments on data from Twitter and Ask.fm, services known to contain high frequencies of bullying. Our experiments indicate that our method can successfully detect new bullying vocabulary. We are currently working on creating a more formal probabilistic model for bullying to robustly incorporate noise and uncertainty.

References

American Academy of Child Adolescent Psychiatry. Facts for families guide. the american academy of child adolescent psychiatry. 2016. URL http://www.aacap.org/AACAP/Families_and_Youth/Facts_for_Families/FFF-Guide/FFF-Guide-Home.aspx.

Bellmore, Amy, Calvin, Angela J., Xu, Jun-Ming, and Zhu, Xiaojin. The five w's of bullying on twitter: Who, what, why, where, and when. *Computers in Human Behavior*, 44:305–314, 2015.

Boyd, Danah. *It's Complicated*. Yale University Press, 2014.

Chen, Ying, Zhou, Yilu, Zhu, Sencun, and Xu, Heng. Detecting offensive language in social media to protect adolescent online safety. *International Conference on Privacy, Security, Risk and Trust (PASSAT), and International Conference on Social Computing (SocialCom)*, pp. 71–80, 2012.

Dinakar, Karthik, Reichart, Roi, and Lieberman, Henry. Modeling the detection of textual cyberbullying. *International Conference on Weblog and Social Media - Social Mobile Web Workshop*, 2011.

ditchthelabel.org. Ditch the label-anti-bullying charity: The annual cyberbullying survey. 2013. URL http://www.ditchthelabel.org/.

Goldman, Russell. Teens indicted after allegedly taunting girl who hanged herself. 2010. URL http://abcnews.go.com/Technology/TheLaw/.

Hosseinmardi, Homa, Han, Richard, Lv, Qin, Mishra, Shivakant, and Ghasemianlangroodi, Amir. Towards understanding cyberbullying behavior in a semi-anonymous social network. *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pp. 244–252, August 2014a.

Hosseinmardi, Homa, Li, Shaosong, Yang, Zhili, Lv, Qin, Rafiq, Rahat Ibn, Han, Richard, and Mishra, Shivakant. A comparison of common users across instagram and ask.fm to better understand cyberbullying. *Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on*, pp. 355–362, 2014b.

Hosseinmardi, Homa, Mattson, Sabrina Arredondo, Rafiq, Rahat Ibn, Han, Richard, Lv, Qin, and Mishra, Shivakant. Detection of cyberbullying incidents on the instagram social network. *Association for the Advancement of Artificial Intelligence*, 2015.

Huang, Qianjia and Singh, Vivek Kumar. Cyber bullying detection using social and textual analysis. *SAM '14 Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, pp. 3–6, 2014.

Mahendiran, Aravindan, Wang, Wei, Arredondo, Jaime, Lira, Sanchez, Huang, Bert, Getoor, Lise, Mares, David, and Ramakrishnan, Naren. Discovering evolving political vocabulary in social media.

Margono, Hendro, Yi, Xun, and Raikundalia, Gitesh K. Mining indonesian cyber bullying patterns in social networks. *Proceedings of the Thirty-Seventh Australasian Computer Science Conference (ACSC 2014)*, 147, January 2014.

Massoudi, Kamran, Tsagkias, Manos, de Rijke, Maarten, and Weerkamp, Wouter. Incorporating query expansion and quality indicators in searching microblog posts. *AIR11. Springer*, 15(5):362–367, November 2011.

Nahar, Vinita, Li, Xue, and Pang, Chaoyi. An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering.*, 3(5):238–247, May 2013.

Ptaszynski, Michal, Dybala, Pawel, Matsuba, Tatsuaki, Masui, Fumito, Rzepka, Rafal, and Araki, Kenji. Machine learning and affect analysis against cyberbullying. *In the 36th AISB*, pp. 7–16), 2010.

Ramakrishnan, Naren, Butler, Patrick, Muthiah, Sathappan, Self, Nathan, Khandpur, Rupinder, and et al. 'Beating the news' with EMBERS: Forecasting civil unrest using open source indicators. *KDD*, 2014.

Reynolds, Kelly, Kontostathis, April, and Edwards, Lynne. Using machine learning to detect cyberbullying. *10th International Conference on Machine Learning and Applications and Workshops (ICMLA)*, 2:241–4, 2011.

Smith-Spark, Laura. Hanna Smith suicide fuels calls for action on ask.fm cyberbullying. 2013. URL http://www.cnn.com/2013/08/07/world/europe/uk-social-media-bullying/.

Wang, Jing, Iannotti, Ronald J., and Nansel, Tonja R. School bullying among us adolescents: Physical, verbal, relational and cyber. *Journal of Adolescent Health*, 45:368–375, 2009.

Yin, Dawei, Xue, Zhenzhen, Hong, Liangjie, Davison, Brian D., Kontostathis, April, and Edwards, Lynne. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB 2.0*, 2009.