PI-Bully: Personalized Cyberbullying Detection with Peer Influence

Lu Cheng1, Jundong Li1, Yasin Silva2 and Deborah Hall3 Huan Liu1
1Computer Science and Engineering, Arizona State University
2Mathematical and Natural Sciences, Arizona State University
3Social and Behavioral Sciences, Arizona State University
{lcheng35,jundongl,ysilva,d.hall,huanliu}@asu.edu

Abstract
Cyberbullying has become one of the most pressing online risks for adolescents and has raised serious concerns in society. Recent years have witnessed a surge in research aimed at developing principled learning models to detect cyberbullying behaviors. These efforts have primarily focused on building a single generic classification model to differentiate bullying content from normal (non-bullying) content among all users. These models treat users equally and overlook idiosyncratic information about users that might facilitate the accurate detection of cyberbullying. In this paper, we propose a personalized cyberbullying detection framework, PI-Bully, that draws on empirical findings from psychology highlighting unique characteristics of victims and bullies and peer influence from like-minded users as predictors of cyberbullying behaviors. Our framework is novel in its ability to model peer influence in a collaborative environment and tailor cyberbullying prediction for each individual user. Extensive experimental evaluations on real-world datasets corroborate the effectiveness of the proposed framework.

1 Introduction
Despite variability in how cyberbullying is defined [Kowalski et al., 2014] and the extent to which it overlaps with or is viewed as being distinct from cyberaggression [Smith, 2012], there is general consensus among researchers that cyberbullying describes the use of electronic forms of communication to intentionally harm or harass others. Cyberbullying has long been one of the most common online risks for adolescents, however, the rapid growth in the use of social media platforms (e.g., Twitter1) has dramatically increased the potential for cyberbullying to occur. The importance of research that increases the accuracy of cyberbullying detection is underscored by the harmful impact of cyberbullying on victims, which can include negative outcomes such as depression, low self-esteem, and suicidal thoughts and behav-
due to the curse of dimensionality. As a result, it is crucial to build a noise-resilient model to alleviate the negative impact of these uninformative features. Second, in spite of considerable diversity in users’ personalities, they also share some common attributes and behaviors. In this regard, it is important to capture the commonality shared by all users as well as idiosyncratic aspects of the personality of each individual for automatic cyberbullying detection. Third, in real-world interactions, victims and perpetrators of cyberbullying are influenced by peers, and the influence from different users can be quite diverse. Hence, developing a way to encode the diversity of peer influence for cyberbullying detection is imperative. The main contributions of this paper are:

- We formally define the problem of personalized cyberbullying detection with peer influence in a collaborative environment. The core idea of our formulation is to customize the prediction for individuals.
- We propose a novel cyberbullying detection framework which consists of three components: (1) a global component that identifies the commonalities among all users; (2) a personalized component that captures the idiosyncratic characteristics of each individual; and (3) a collaborative/peer influence component that can quantify the diverse influence from other users.
- We perform empirical experiments on multiple real-world datasets from microblogging platforms to corroborate the efficacy of the proposed framework.

3 Problem Statement

Suppose \( U \) users generate \( N \) posts in a social media platform. Let \( \{(x_j^i, y_j^i) \mid j = 1, \ldots, N_i\} \) represent the posts from the \( i \)-th user and define \( N = \sum_{i=1}^{U} N_i \). Each specific post \( j \) from user \( i \) is represented by \( (x_j^i, y_j^i) \). \( x_j^i \in D \) represents the post’s features, \( D \) is the number of features, and \( y_j^i \) denotes the class label associated with this post. In this work, we assume each post is associated with two possible labels \( y_j^i \in \{0, 1\} \), where \( y_j^i = 1 \) denotes that the post is a cyberbullying message and \( y_j^i = 0 \) otherwise. Then \( X = [x_1^1; \ldots; x_{N_1}^1; \ldots; x_1^U; \ldots; x_{N_U}^U] \in R^{N \times D} \) is the feature representation of all these \( N \) posts and \( y = [y_1^1; \ldots; y_{N_1}^1; \ldots; y_1^U; \ldots; y_{N_U}^U] \in \{0, 1\}^N \) is the corresponding label vector. With the aforementioned notation, we now formally define the problem of personalized cyberbullying detection with peer influence in a collaborative environment as follows:

Given the feature representation \( X \) of \( N \) social media posts from \( U \) users and the label vector \( y \) of these \( N \) posts, the goal is to train a binary classification model to predict the labels of online social media posts (bullying or normal). In particular, during the learning phase, we would like to (1) tailor the prediction for each user by capturing commonalities among multiple users and individual characteristics; and (2) quantify the way each user is influenced by like-minded users.

3 The Proposed Framework

In this section, we describe how to build a generic classification model to identify cyberbullying behaviors when the feature representation of users’ posts are sparse, noisy, and high-dimensional. We first show a global model that is designed to capture the commonality shared by all users and then describe the mechanisms to model users’ idiosyncrasies. In addition, as the occurrence of cyberbullying is heavily related to peer influence, we investigate how to quantify the influence from like-minded users so that personalized modeling can benefit from users with similar behavior. Finally, we show how to predict an unlabeled post from an unseen user using the PI-Bully model and briefly introduce the optimization algorithm and its complexity. Fig. 1 illustrates the overview architecture of the PI-Bully framework.

3.1 Building the Personalized Model

Previous efforts in cyberbullying detection have been primarily devoted to the development of a global classification model to capture the commonalities among users. It formulates cyberbullying detection as a binary classification task:

\[
\min_w \sum_{i=1}^{U} \sum_{j=1}^{N_i} f(x_j^i, y_j^i, w) + \lambda_1 \|w\|_1, \tag{1}
\]

where \( w \in D \) is the global classification model that applies to all users (\( U \)) and \( \lambda_1 \) controls the sparsity of the model. As feature representation in the traditional way may lead to an extremely high-dimensional representation of posts, we integrate feature selection [Li et al., 2017] into the model by imposing an \( \ell_1 \)-norm sparse regularization term. \( f(\cdot) \) is a loss function to measure the loss between the ground truth class labels and predicted class labels. In this work, we use the squared loss function, i.e., \( f(x_j^i, y_j^i, w) = (w^Tx_j^i - y_j^i)^2 \), but the model could also use other functions such as hinge loss and cross entropy loss.

In spite of the empirical success of global classification models, research advances in psychology indicate that cyberbullying is correlated with a number of individual features—such as personality traits (e.g., [Baughman et al., 2012]), attitudes and beliefs (e.g., [Hinduja and Patchin, 2013]), and motives (e.g., [Gradinger et al., 2011])—that vary from user to user. In short, although users may share a number of inherent characteristics, they are also highly idiosyncratic. To this end, we assume each user \( u_i \) has a personalized model \( M_i \in D \) in addition to the global model \( w \in D \). Moreover, we impose an \( \ell_1 \)-norm sparse regularization term on each personalized model \( M_i \) to alleviate the curse of dimensionality. Hence, we obtain the following optimization framework:

\[
\min_{w,M_i} \sum_{i=1}^{U} \sum_{j=1}^{N_i} f(x_j^i, y_j^i, w + M_i) + \lambda_1 (\|w\|_1 + \sum_{i=1}^{U} \|M_i\|_1). \tag{2}
\]

3.2 Characterizing Peer Influence

The process of parameter learning in the above personalized model can be problematic due to the limited amount of training data for each user. Because of this, the generated personalized model can easily suffer from overfitting and have poor generalization ability on unseen test data. To address this problem, we decompose the personalized model \( M_i \) of
each user into a personalized component, \( P_i \in \mathbb{R}^D \), which encodes a user’s inherent traits, and a collaborative/peer influence component, \( Q_i \in \mathbb{R}^D \). The goal for including this collaborative/peer influence component is to extrapolate information about the way a user experiences cyberbullying by considering the experiences of similar users. By doing this, we aim to capture the influence of similar, like-minded users in the way a person experiences cyberbullying. The collaborative/peer influence component \( Q_i \) is customized for each user and is estimated by a weighted average of the personalized component \( P_i \) of other users. The integration of this component is motivated by empirical findings in psychology indicating similarity within child and adolescent peer groups in both bullying behavior and victimization [Espelage et al., 2003]. The objective function in Eq. (2) can then be reformulated as follows:

\[
\min_{w, P_i, Q_i} \sum_{i=1}^{U} \sum_{j=1}^{N_i} f(x_{ij}, y_{ij}, w + P_i + Q_i) + \\
\lambda_1(\|w\|_1 + \|P\|_1) + \lambda_2 \sum_{i=1}^{U} \|Q_i\| - \sum_{j=1}^{U} s_{ij}P_j^2, \tag{3}
\]

where \( \lambda_2 \) balances the contribution of collaborative/peer influence for personalized cyberbullying detection and \( s_{ij} \in S \) denotes how user \( u_i \) is influenced by user \( u_j \) (the influence from different peers could vary significantly). Here, we provide a more intuitive illustration of these components: As the proposed model focuses mainly on text, \( W \) captures the common language used globally across all the users, \( P \) represents the unique language characteristics of individual users, and \( Q \) is a parameter that captures additional predictive value that can be drawn from between-user language similarities.

In this work, we exploit the method presented in [Anava and Levy, 2016] to adaptively learn the optimal neighborhood structure (i.e., the most influential neighbors) around each user to quantify the diversity of peer influence. Specifically, we leverage the \( k^* \)-NN algorithm in [Anava and Levy, 2016] to process the data matrix \( X \) and generate the User-to-User peer influence matrix \( S \) with dimensions \( U \times U \). The \( i \)-th row in \( S \) represents the pairwise similarities between user \( i \) and other users. In Eq. (3), we can also observe that the personalized models \( P_i \) of similar users are explicitly correlated with each other through the last term (collaborative/peer influence component), which implicitly generates additional data for each user to train the personalized model. In summary, from Eq.(3), we can observe that the PI-Bully model for each user \( u_i \) has three components: (1) a global model \( w \) that captures the shared traits of all users; (2) a personalized model \( P_i \) that captures the unique characteristics of the user; and (3) a collaborative/peer influence component that quantifies the way the user is influenced by like-minded users.

### 3.3 Inference on Unlabeled Data

We describe next how the PI-Bully framework predicts whether an unlabeled post \( x \in \mathbb{R}^D \) is a cyberbullying message or not, given the learned parameters \( w, P, Q \).

There are two cases to discuss. If the user \( u \) of the post \( x \) appears in the training dataset, we can directly use the learned model parameters to make the prediction. In this case, the classifier for the new post \( x \) is \( c = w + P_u + Q_u \). If the post is from a new user \( m \) that does not appear in the training dataset, we first leverage the same mechanism as the one used in [Anava and Levy, 2016] to measure the influence of existing users in the training dataset on user \( m \). Then, we estimate the model parameters for this new user by solving the following optimization problem (a.k.a. Weber problem [Hallac et al., 2015]):

\[
\min_{P_m, Q_m} \sum_{j=1}^{U} s_{im}(\|P_m - P_i\|_2 + \|Q_m - Q_i\|_2), \tag{4}
\]

where \( s_{im} \) indicates how the new user \( m \) is influenced by user \( i \). After solving the above Weber problem, the classifier for the new post is specified as \( c = w + P_m + Q_m \). Using the classifier \( c \) for the new unlabeled post, we predict that the new post is cyberbullying if \( x \cdot c \geq 0.5 \), and normal otherwise.
3.4 Optimization

The proposed PI-Bully model in Eq. (3) has three sets of model parameters: $w$, $P_i$ ($i = 1, ..., U$) and $Q_i$ ($i = 1, ..., U$). We can observe that the objective function is not convex regarding these three sets of parameters simultaneously. In addition, it is not smooth as well due to the $l_1$-norm sparse regularization terms. We address these problems using the Alternating Direction Method of Multipliers (ADMM) [Boyd et al., 2011] approach and adapting advances of fast convergence rate from Fast Iterative Shrinkage-Thresholding (FISTA) [Beck and Teboulle, 2009] for the local optimal solutions. Details are omitted here due to the space constraint.

4 Experimental Evaluation

In this section, we present experimental results to evaluate the effectiveness of the proposed PI-Bully model. In particular, we aim to answer the following research questions: (1) Can the proposed framework achieve better cyberbullying detection performance than existing models? (2) What is the impact of the different components of the proposed PI-Bully framework? and (3) How robust is the proposed model w.r.t. the different model parameters? All of the experiments were executed on MATLAB 2017a using a Mac OS X system with Intel Core i5 and 8GB of RAM. We report averaged evaluation scores based on multiple runs of the experiments.

4.1 Datasets

We used two real-world datasets crawled from the microblogging platform, Twitter. The first dataset3 (referred to Xu et al.), with 3095 data samples, was published in [Xu et al., 2012]. Note that the original dataset consisted of 7321 tweets, among which only 3095 tweets were publicly available at the time we crawled using the Tweet IDs. Following the procedure suggested by [Nand et al., 2016], we collected the second dataset (referred as Authors) via the Twitter streaming API from September 19th to 25th (2017) using the following keywords: nerd, gay, loser, freak, emo, whale, pig, fat, wannabe, poser, whore, should, die, slept, caught, suck, slut, live, afraid, fight, pussy, cunt, kill, dick, bitch.

Early in the experimental design phase, a decision was made to maximize the labeling quality (while maintaining a proper dataset size) instead of prioritizing only the data size. Following this guideline, we extracted 20000 tweets to be manually labeled by well-trained human annotators (with backgrounds in psychology and computer science) over a period of two months. The human annotators followed coding guidelines that were similar to the ones described in [Nand et al., 2016]. Each tweet was initially labeled by two annotators and the agreement level between the two annotators in this stage was 80%. A third annotator was asked to resolve the conflicts identified in the initial annotation phase.

After conflict resolution and data cleaning, we finally obtained the Authors dataset with a total number of 19994 labeled tweets. Table 1 shows the statistics for the two datasets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th># Users</th>
<th># Tweets</th>
<th># Bullying</th>
<th># Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xu et al.</td>
<td>2,948</td>
<td>3,095</td>
<td>1,794</td>
<td>1,301</td>
</tr>
<tr>
<td>Authors</td>
<td>9,833</td>
<td>19,994</td>
<td>3,845</td>
<td>16,149</td>
</tr>
</tbody>
</table>

It is important to note that the proportions of bullying to normal messages in the two datasets are different. Whereas 57.96% of the posts in the Xu et al. dataset are bullying messages, 19.23% of the posts in the Authors dataset are bullying interactions. This latter percentage is similar to the one found in [Hosseinmardi et al., 2015] for Instagram data and more closely represents the proportion of bullying to non-bullying messages in the real-world. Our dataset can be downloaded from http://www.public.asu.edu/ lcheng35/.

We performed psychometric analysis to obtain features for each tweet in the aforementioned datasets through Linguistic Inquiry Word Count (LIWC) [Pennebaker et al., 2001]. Specifically, LIWC counts words that belong to certain categories in psychology. For example, the word “cry” belongs to five categories: sadness, negative emotion, overall affect, verb, and past tense verb. The results of previous research show that such psychometric analysis can improve the performance of cyberbullying detection [Nand et al., 2016].

4.2 Performance Evaluation

To answer the first question, we compare PI-Bully with common text classification models ($k$NN, Random Forest, Linear SVM, and Logistic Regression) and two text-based cyberbullying detection models (Bully [Xu et al., 2012] and SICD [Dani et al., 2017]). We specify these models below.

- $k$NN: It predicts the class labels of unlabeled instances using a $k$-nearest classifier where the distance metric is specified as the Euclidean distance.
- Random Forest (RF): It builds a random forest model on the training data with all of the contextual features.
- Linear SVM (SM): It implements a regularized linear support vector machine model with stochastic gradient descent (SGD) learning.
- Logistic Regression (LR): It is a widely used statistical model that, in its basic form, uses a logistic function to model a binary dependent variable.
- Bully [Xu et al., 2012]: This model extracts several NLP features including unigrams, unigrams+bigrams, and POS colored N-grams to train a SVM model.
- SICD [Dani et al., 2017]: It uses both content (i.e., TF-IDF) and sentiment information embedded in the user-generated content to boost the performance of cyberbullying detection.

Our evaluation methods include several widely-used metrics - AUC, Precision, Recall, and F1 scores. The primary difference between the Accuracy, F1, and AUC scores is that the AUC score considers all possible thresholds for classification (e.g., 0.5 for binary label \{0, 1\}) and is a more appropriate metric as the cyberbullying datasets are typically imbalanced, i.e., each class does not make up an equal proportion of the dataset. Imbalanced datasets may affect the trade-off

---

3https://twitter.com/

1http://research.cs.wisc.edu/bullying/data.html
between recall and precision. In the context of cyberbullying detection, missing a positive instance is usually less desirable than incorrectly labeling a negative instance. Hence, achieving high recall scores is particularly important.

We use 80% of the datasets for training and the rest for testing, and present the results in Tables 2-3. The best and the second best scores are highlighted with bold and underscored text, respectively. We can observe that for both datasets, PI-Bully achieves the best Recall, F1, and AUC scores while most baseline models present poor Recall and AUC scores. This is especially apparent for Authors dataset, which more closely represents the proportion of bullying to normal messages in the real-world. We can conclude that PI-Bully significantly boosts the classification of positive samples and leads to the improved overall performance of cyberbullying detection. Results of a pairwise Wilcoxon signed-rank test indicate that the improvement of PI-Bully is significant, with a 0.05 significance level.

### 4.3 Impact of Different Model Components

For each user, PI-Bully is composed of three components: (1) the global model w common to all users; (2) a personalized component PI-i that is customized for each individual user; and (3) a collaborative/peer influence component that quantifies the influence from like-minded users. We compare in this subsection the following variants:

- The personalized component (P): a variant of the proposed PI-Bully framework that only includes the personalized component PI-i for each user.
- The global model (G): a variant that only includes the global model w.
- Global+Personalized (G+P): a variant of PI-Bully without the peer influence component.
- Global+Influence (G+I): a variant of PI-Bully that eliminates the personalized component PI-i for each user.

We compare these four variants with the proposed PI-Bully model. For these experiments, the percentage of training data is incrementally increased from 20% to 80%. The comparison of these four variants and the proposed PI-Bully model is illustrated in Fig. 2. We highlight the following key findings:

- The personalized component P is inferior to the global model G. The main reason is that P often suffers from the over-parameterization issue due to the lack of training data, whereas the global model G can collect more data to capture the commonalities among all users during the training phase.
- Both the G+I model and the G+P model outperform the global model G. These results validate the importance of incorporating the personalized and peer influence components for personalized cyberbullying detection.
- The proposed PI-Bully framework achieves the best performance and the dominance tends to become more obvious as the dataset becomes more imbalanced, i.e. the Authors dataset. This shows the benefits of considering the three proposed components.

### 4.4 Parameter Analysis

The PI-Bully model has two key parameters: λ1 controls the balance between the personalized and common features in the model learning phase and λ2 regulates the importance of peer influence in PI-Bully. To investigate the effect of these two parameters, we fix one parameter at a time and vary the other one to evaluate how it affects the classification performance. We vary the values of λ1 and λ2 among \{1e-7,1e-5,1e-3,0.1,10\} and show the AUC and F1 scores in Fig. 3.  

---

**Table 2: Performance comparison w.r.t. Authors dataset.**

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN</td>
<td>0.663</td>
<td>0.517</td>
<td>0.581</td>
<td>0.622</td>
</tr>
<tr>
<td>SVM</td>
<td>0.685</td>
<td>0.646</td>
<td>0.665</td>
<td>0.714</td>
</tr>
<tr>
<td>RF</td>
<td>0.708</td>
<td>0.544</td>
<td>0.605</td>
<td>0.678</td>
</tr>
<tr>
<td>LR</td>
<td>0.680</td>
<td>0.646</td>
<td>0.663</td>
<td>0.711</td>
</tr>
<tr>
<td>Bully</td>
<td>0.653</td>
<td>0.508</td>
<td>0.571</td>
<td>0.709</td>
</tr>
<tr>
<td>SCD</td>
<td>0.803</td>
<td>0.763</td>
<td>0.696</td>
<td>0.791</td>
</tr>
<tr>
<td>PI-Bully</td>
<td>0.425</td>
<td>0.887</td>
<td>0.574</td>
<td>0.844</td>
</tr>
</tbody>
</table>

**Table 3: Performance comparison w.r.t. Xu et al. dataset.**

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN</td>
<td>0.663</td>
<td>0.517</td>
<td>0.581</td>
<td>0.622</td>
</tr>
<tr>
<td>SVM</td>
<td>0.685</td>
<td>0.646</td>
<td>0.665</td>
<td>0.714</td>
</tr>
<tr>
<td>RF</td>
<td>0.708</td>
<td>0.544</td>
<td>0.605</td>
<td>0.678</td>
</tr>
<tr>
<td>LR</td>
<td>0.680</td>
<td>0.646</td>
<td>0.663</td>
<td>0.711</td>
</tr>
<tr>
<td>Bully</td>
<td>0.653</td>
<td>0.508</td>
<td>0.571</td>
<td>0.709</td>
</tr>
<tr>
<td>SCD</td>
<td>0.727</td>
<td>0.609</td>
<td>0.663</td>
<td>0.722</td>
</tr>
<tr>
<td>PI-Bully</td>
<td>0.656</td>
<td>0.740</td>
<td>0.695</td>
<td>0.802</td>
</tr>
</tbody>
</table>
Whereas large $\lambda_1$ values that overemphasize the personalized features can result in relatively poor performance w.r.t. AUC, they can slightly improve the F1 score. Consequently, a proper value of $\lambda_1$ enables the identification of the most representative personalized features for both positive and negative samples. The proposed framework is more robust to changes of $\lambda_2$, and presents an increasing trend when $\lambda_1$ is in a certain range. We select $\lambda_2$ based on cross-validations for all of the previous experiments. In summary, PI-Bully is not significantly sensitive to any of the two parameters over a wide range of values, and consequently can be tuned for various application purposes.

5 Related Work

Cyberbullying is a serious issue with significant negative societal consequences. To date, a number of dedicated learning algorithms have been proposed to identify cyberbullying instances. Most existing methods adopt a two-stage approach to detect cyberbullying: they first apply feature engineering to identifying feature sets that enable capturing cyberbullying patterns and then employ off-the-shelf machine learning classifiers to detect cyberbullying behaviors. Typically, the feature set includes text-based [Xu et al., 2012; Dani et al., 2017; Bellmore et al., 2015] and network-based features [Squicciarini et al., 2015; Al-garadi et al., 2016]. Various methods differ in the types of features used for classification. For example, Dinakar et al. [Dinakar et al., 2011] concatenated TF-IDF features, POS tags of frequent bigrams, and profane words as content features to detect cyberbullying behaviors. Xu et al. [Xu et al., 2012] presented several off-the-shelf tools such as Bag-of-Words models and LSA- and LDA-based representation learning to predict bullying traces in Twitter. In [Dadvar et al., 2013], the authors made use of gender-specific features and contextual features, such as users’ previous posts and the use of profane words, to improve the performance of cyberbullying detection. Dani et al. [Dani et al., 2017] proposed the SICD model which incorporates sentiment into content features. Their goal was to facilitate cyberbullying detection by capturing the sentiment consistency of normal and bullying posts. Bellmore et al. [Bellmore et al., 2015] used a dictionary including words in a Twitter corpus to construct a frequency vector for each tweet and trained a text classifier to answer core questions about cyberbullying (“Who, What, Why, Where, and When”).

With the increasing prevalence of social networking systems, network-based features e.g., the number of friends, network embeddedness, and relational centrality are also used to detect cyberbullying behaviors [Squicciarini et al., 2015]. For instance, previous work by [Al-garadi et al., 2016] studied a model that integrated the use of activity information, user information, and tweet content features. Cyberbullying has also been studied in other social media platforms such as Ask.fm [Li et al., 2014], Instagram [Hosseinmardi et al., 2015; Cheng et al., 2019b; Cheng et al., 2019a], and Vine [Rafiq et al., 2016]. In addition, some work has focused on developing systems and applications to identify cyberbullying behaviors on social network platforms [Silva et al., 2016a; Silva et al., 2016b]. The authors aim to estimate the probability of an individual experiencing cyberbullying considering the received messages and various cyberbullying risk factors.

6 Conclusions and Future Work

Existing efforts toward detecting cyberbullying have focused heavily on building generic classification models for all users that seek to distinguish bullying behaviors from normal content. These methods, however, ignore unique characteristics that are embedded in the user-generated content. Empirical findings from psychology highlight the role of individual differences—reflected in users’ unique personality traits, attitudes, motives, etc.—and influence from like-minded users as predictors of cyberbullying. In this paper, we propose a principled personalized cyberbullying detection framework, PI-Bully, that draws on these interdisciplinary findings to tailor and improve the prediction of cyberbullying behaviors.

Future work in cyberbullying detection can be performed in several key areas. First, there has been limited research examining predictive models that take temporal properties and patterns of cyberbullying into account, which stands to contribute to a deeper understanding of the nature of cyberbullying across research disciplines. Second, there is a growing need for cyberbullying detection models that rely on limited or aggregated data—in part, due to the difficulty of accessing social media data. This underscores the need for models that can achieve high accuracy while relying on limited, incomplete, anonymized, or aggregated data. Finally, future work should be aimed at better integrating interdisciplinary empirical findings—such as those from psychology and related social sciences—into computational models that detect cyberbullying. Interdisciplinary synergies hold particular promise for addressing and preventing this major social problem.

Acknowledgements

This material is based upon work supported by the National Science Foundation (NSF) Grant 1719722.

References


