PI-Bully: Personalized Cyberbullying Detection with Peer Influence

Lu Cheng[1] , **Jundong Li**[1] , **Yasin Silva**[2] , **Deborah Hall**[3] , **Huan Liu**[1]

[1]Computer Science and Engineering, Arizona State University
[2]Mathematical and Natural Sciences, Arizona State University
[3]Social 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 the 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 a general consensus 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., Twitter[1]) 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 behaviors [Xu *et al.*, 2012]. Within computer science, existing efforts toward detecting cyberbullying behaviors have primarily focused on building a generic binary classification model for all users by analyzing user-generated content [Xu *et al.*, 2012; Dani *et al.*, 2017]. Although these approaches have yielded satisfactory detection performance in practice, these models treat all users equally and, thus fail to capture aspects of cyberbullying that are shaped by the unique characteristics and circumstances of each individual.

Empirical findings within psychology indicate that individual difference variables – that is, characteristics that make people different from one another, such as personality traits – are important predictors of computer-mediated behaviors [Kowalski *et al.*, 2014; Goodboy and Martin, 2015]. For example, three personality traits, referred to collectively as the *Dark Triad*, are correlated with cyberbullying perpetration [Goodboy and Martin, 2015]. Specifically, individuals who exhibit higher levels of (i) machiavellianism (i.e., a desire to manipulate others), (ii) psychopathy (i.e., low levels of empathy), and (iii) narcissism (i.e., feelings of superiority relative to others) are more likely to bully others online [Goodboy and Martin, 2015]. Broadly, these findings highlight the strong potential for models that characterize and take into account the unique attributes of cyberbullies and their victims to facilitate a better understanding of cyberbullying behaviors. Previous empirical work in psychology has also identified patterns of similarity in bullying behaviors and victimization within the child and adolescent peer groups, which likely result from peer influence within established groups, as well as self-selection of youth into friendship groups with similar others [Espelage *et al.*, 2003]. Therefore, given a target user, a key research question is how to model his/her idiosyncratic characteristics and quantify the peer influence from similar users to facilitate cyberbullying detection.

In this paper, we study the novel problem of personalized cyberbullying detection with peer influence in a collaborative environment. Notably, we use the term "personality" broadly to refer to users' unique collection of traits, characteristics, and circumstances. Building a personalized cyberbullying detection framework that is customized to each individual presents multiple challenges. First, users' information in social media platforms is often very noisy, containing irrelevant and redundant features that may jeopardize the learning performance. As a result, it is crucial to building a noise-resilient model to alleviate the negative impact of these uninformative

[1]https://twitter.com/

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 as follows:

- 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.

The remainder of this paper is organized as follows. In Section 2, we formally define the problem of personalized cyberbullying detection with peer influence in a collaborative environment. Section 3 describes the proposed PI-Bully framework in detail. In Section 4, we discuss the used datasets, the experimental settings, and our findings from experiments. Section 6 reviews related work and Section 7 concludes the paper and describes some paths for future work.

2 Problem Statement

Suppose U users generate N posts in a social media platform. Let $\{(\mathbf{x}_j^i, y_j^i) \,|\, j = 1, ..., N_i\}$ 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 $(\mathbf{x}_j^i, y_j^i)$. $\mathbf{x}_j^i \in \mathbb{R}^D$ represents the post's features, D is the number of features, and y_j^i denotes the class label associated with the 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 $\mathbf{X} = [(\mathbf{x}_1^1)^T; ...; (\mathbf{x}_{N_1}^1)^T; ...; (\mathbf{x}_1^U)^T; ...; (\mathbf{x}_{N_U}^U)^T] \in \mathbb{R}^{N \times D}$ is the feature representation of all these N posts and $\mathbf{y} = [y_1^1; ...; y_{N_1}^1; ...; y_1^U; ...y_{N_U}^U] \in \{0, 1\}^N$ is the corresponding label vector. With the aforementioned notations, we define the problem of *personalized cyberbullying detection with peer influence in a collaborative environment* as follows:

Given the feature representation $\mathbf{X}$ of N social media posts from U users and the label vector $\mathbf{y}$ 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. 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 such that personalized modeling can benefit from users with similar behaviors. 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 time complexity. Fig. 1 illustrates the overview workflow 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:

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \sum_{j=1}^{N_i} f(\mathbf{x}_j^i, y_j^i, \mathbf{w}) + \lambda_1 \|\mathbf{w}\|_1, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^D$ is the global classification model that applies to all users. As vast majority of feature representation methods for social media posts may lead to the inclusion of uninformative features, we integrate feature selection [Li *et al.*, 2017] into the classifier by imposing an ℓ_1-norm sparse regularization term, where λ_1 controls the sparsity of the model. $f(\cdot)$ 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., $f(\mathbf{x}_j^i, y_j^i, \mathbf{w}) = (\mathbf{w}^T \mathbf{x}_j^i - y_j^i)^2$, 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 $\mathbf{M}_i \in \mathbb{R}^D$ in addition to the global model $\mathbf{w} \in \mathbb{R}^D$. Moreover, we impose an ℓ_1-norm sparse regularization term on each personalized model $\mathbf{M}_i$ to reduce the model complexity. Hence, we obtain the following optimization framework:

$$\min_{\mathbf{w}, \mathbf{M}_i} \sum_{i=1}^{U} \sum_{j=1}^{N_i} f(\mathbf{x}_j^i, y_j^i, \mathbf{w} + \mathbf{M}_i) + \lambda_1 (\|\mathbf{w}\|_1 + \sum_{i=1}^{U} \|\mathbf{M}_i\|_1). \tag{2}$$

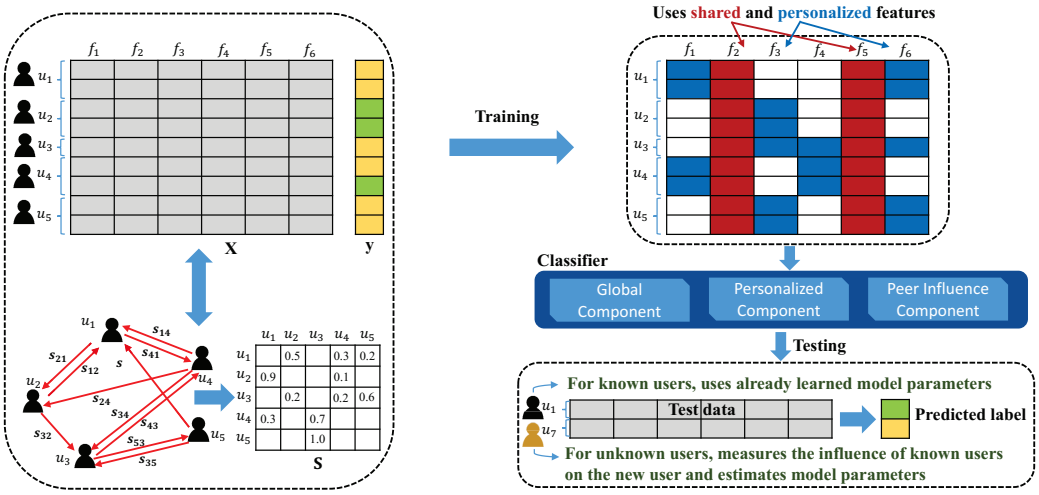

Figure 1: The proposed PI-Bully framework. We first leverage the data matrix $\mathbf{X}$ to compute the influence matrix $\mathbf{S}$, which quantifies how users could influence each other. For example, u_2 is influenced more by u_1 than u_4 as $s_{21} = 0.9$ is larger than $s_{24} = 0.1$. Then, in the training phase, we use both shared and user-specific features to train a classifier by capturing the commonalities of all users and the idiosyncrasies of each specific user. Finally, in the testing phase, given a set of unlabeled test data, we predict if the new posts are cyberbullying or not.

3.2 Characterizing Peer Influence

The process of model 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 $\mathbf{M}_i$ of each user into a personalized component, $\mathbf{P}_i \in \mathbb{R}^D$, which encodes a user's inherent traits, and a collaborative/peer influence component, $\mathbf{Q}_i \in \mathbb{R}^D$. The goal for including this collaborative/peer influence component is to extrapolate information about the way a user experiences cyberbullying from 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 $\mathbf{Q}_i$ is customized for each user and is estimated by a weighted average of the personalized component $\mathbf{P}_i$ 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 behaviors and victimization [Espelage *et al.*, 2003]. Then the objective function in Eq. (2) can be reformulated as

$$\min_{\mathbf{w},\mathbf{P}_i,\mathbf{Q}_i} \sum_{i=1}^{U}\sum_{j=1}^{N_i} f(\mathbf{x}_j^i, y_j^i, \mathbf{w}+\mathbf{P}_i+\mathbf{Q}_i) + \lambda_1(\|\mathbf{w}\|_1 + \|\mathbf{P}\|_1) + \lambda_2 \sum_{i=1}^{U} \|\mathbf{Q}_i - \sum_{j=1}^{U} s_{ji}\mathbf{P}_j\|_2^2, \tag{3}$$

where $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{U\times D}$ respectively denote the concatenation of personalized component and collaborative component of all users, λ_2 balances the contribution of collaborative/peer influence for personalized cyberbullying detection and $s_{ji} \in \mathbf{S}$ 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 data, $\mathbf{w}$ captures the common language used globally across all the users, $\mathbf{P}$ represents the unique language characteristics of individual users, and $\mathbf{Q}$ is a model 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 $\mathbf{X}$ and generate the User-to-User peer influence matrix $\mathbf{S}$ of size $U \times U$. The i-th row in $\mathbf{S}$ represents the pairwise similarities between user i and other users. In Eq. (3), we can also observe that the personalized models $\mathbf{P}_i$ 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, we can observe that the PI-Bully model for each user u_i has three components: (1) a global model $\mathbf{w}$ that captures the shared traits of all users; (2) a personalized model $\mathbf{P}_i$ that captures the unique characteristics of the user; and (3) a collaborative/peer influence component that quantifies how the user is influenced by like-minded users.

3.3 Inference on Unlabeled Data

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

There are two cases to discuss. If the user u of the post $\mathbf{x}$ 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 $\mathbf{x}$ is $\mathbf{c} = \mathbf{w} + \mathbf{P}_u + \mathbf{Q}_u$. 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]):

$$\min_{\mathbf{P}_m, \mathbf{Q}_m} \sum_{i=1}^{U} s_{im}(\|\mathbf{P}_m - \mathbf{P}_i\|_2 + \|\mathbf{Q}_m - \mathbf{Q}_i\|_2), \quad (4)$$

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 $\mathbf{c} = \mathbf{w} + \mathbf{P}_m + \mathbf{Q}_m$. Using the classifier $\mathbf{c}$ for the new unlabeled post, we predict that the new post is cyberbullying if $\mathbf{x}^T\mathbf{c} \geq 0.5$, and normal otherwise.

3.4 Optimization

The proposed PI-Bully model in Eq. (3) has three sets of model parameters: $\mathbf{w}$, $\mathbf{P}_i$ $(i = 1, ..., U)$ and $\mathbf{Q}_i$ $(i = 1, ..., U)$. 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. Motivated by [Wu and Huang, 2016], we address these problems by using the Alternating Direction Method of Multipliers (ADMM) [Boyd *et al.*, 2011] and Fast Iterative Shrinkage-Thresholding (FISTA) [Beck and Teboulle, 2009] to achieve a local optimal solution. 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. different model hyperparameters?

4.1 Datasets

We use two real-world datasets crawled from the microblogging platform, Twitter[2]. The first dataset[3] (referred to *Xu et.al*), with 3,095 social posts, was published in [Xu *et al.*, 2012]. Note that the original dataset consists of 7,321 tweets, among which only 3,095 tweets were publicly available at the time we crawled using the Tweet IDs. Following the procedure suggested by [Nand *et al.*, 2016], we collect 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 20,000 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.

Datasets	# Users	# Tweets	# Bullying	# Normal
Xu et al.	2,948	3,095	1,794	1,301
Authors	9,833	19,994	3,845	16,149

Table 1: Basic statistics of the two used datasets.

After conflict resolution and data cleaning, we finally obtained the *Authors* dataset with a total number of 19,994 labeled tweets. Table 1 shows the statistics for these two datasets. It is important to note that the proportions of bullying to normal messages in these 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 perform 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 (kNN, *Random Forest*, *Linear SVM*, and *Logistic Regression*) with the same input features and two text-based cyberbullying detection models (*Bully* [Xu *et al.*, 2012] and *SICD* [Dani *et al.*, 2017]). We specify these models below.

- *kNN*: 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 is an ensemble learning method that constructs a multitude of decision trees during training and output the mode of the classes at testing.
- *Linear SVM (SVM)*: It implements a regularized linear support vector machine model with stochastic gradient descent (SGD) learning.
- *Logistic Regression (LR)*: It is an extension of linear regression model for the classification problem with the logistic function as the loss function.
- *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.

[2]https://twitter.com/

[3]http://research.cs.wisc.edu/bullying/data.html

Metrics	Precision	Recall	F1	AUC
kNN	0.663	0.364	0.470	0.652
SVM	0.699	0.469	0.562	0.701
RF	0.708	0.478	0.571	0.707
LR	0.680	0.485	0.566	0.705
Bully	0.653	0.508	0.571	0.709
SICD	**0.803**	0.263	0.396	0.791
PI-Bully	0.425	**0.887**	**0.574**	**0.844**

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

Metrics	Precision	Recall	F1	AUC
kNN	0.663	0.517	0.581	0.662
SVM	0.685	0.646	0.665	0.714
RF	0.681	0.544	0.605	0.678
LR	0.680	0.646	0.663	0.711
Bully	0.708	0.646	0.676	0.725
SICD	**0.727**	0.609	0.663	0.722
PI-Bully	0.656	**0.740**	**0.695**	**0.802**

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

- *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 - Precision, Recall, F1 score, and AUC. The main reason we choose F1 score and AUC rather than Accuracy is that the cyberbullying datasets are typically imbalanced, i.e., each class does not make up an equal proportion of the dataset. Meanwhile, imbalanced datasets may affect the trade-off 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 is particularly important.

In the experiments, we use 80% of the datasets for training and the rest for testing, the averaged classification results based on ten runs are shown in Tables 2-3. We select the hyperparameters based on cross-validation on the training data. A detailed hyperparameter analysis can be seen in Sec. 4.4. 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 $\mathbf{w}$ common to all users; (2) a personalized component $\mathbf{P}_i$ 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:

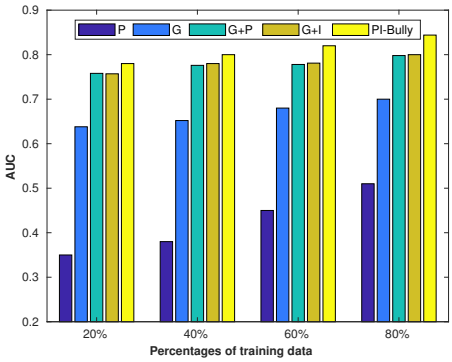

Figure 2: Performance evaluation of different components on the *Authors* dataset.

- The personalized component (P): a variant of the proposed PI-Bully framework that only includes the personalized component $\mathbf{P}_i$ for each user.
- The global model (G): a variant that only includes the global model $\mathbf{w}$.
- 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 $\mathbf{P}_i$ for each user.

We compare these four variants with the proposed PI-Bully model using the *Authors* dataset. 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 Hyperparameter Analysis

The PI-Bully model has two key hyperparameters: λ_1 controls the complexity of the model (i.e., it balances the sparsity of personalized features and common features in the model learning phase) while λ_2 regulates the importance of peer influence in PI-Bully. To investigate the effect of these two

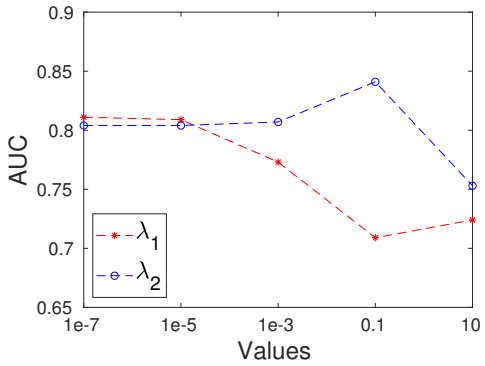

Figure 3: Effects of λ_1, λ_2 w.r.t. AUC on the *Authors* dataset.

parameters, we fix one parameter at a time (λ_1=1e-7,λ_2=1e-7 respectively) and vary the other one to evaluate how it affects the classification performance. We vary the values of λ_1 and λ_2 among {1e-7,1e-5,1e-3,0.1,10} and show the AUC score in Fig. 3. We can see the performance keeps stable when λ_1 is specified as a small number, while large λ_1 overemphasizes the personalized features can result in relatively poor performance. Consequently, a proper value of λ_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 λ_2, and presents an increasing trend when λ_2 is in a certain range. In summary, the performance of PI-Bully is relatively stable when the hyperparameters are varied in a certain range, and thus 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.*, 2019], 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; 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 heavily focused 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 difference variables – 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.

Acknowledgements

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

References

[Al-garadi *et al.*, 2016] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior*, 63:433–443, 2016.

[Anava and Levy, 2016] Oren Anava and Kfir Levy. k*-nearest neighbors: From global to local. In *NIPS*, pages 4916–4924, 2016.

[Baughman *et al.*, 2012] Holly M Baughman, Sylvia Dearing, Erica Giammarco, and Philip A Vernon. Relationships between bullying behaviours and the dark triad: A study with adults. *Personality and Individual Differences*, 52(5):571–575, 2012.

[Beck and Teboulle, 2009] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIIMS*, 2(1):183–202, 2009.

[Bellmore *et al.*, 2015] Amy Bellmore, Angela J Calvin, Jun-Ming Xu, and Xiaojin Zhu. The five w's of "bullying" on twitter: who, what, why, where, and when. *Computers in human behavior*, 44:305–314, 2015.

[Boyd *et al.*, 2011] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[Cheng *et al.*, 2019] Lu Cheng, Jundong Li, Yasin Silva, Deborah Hall, and Huan Liu. Xbully: Cyberbullying detection within a multi-modal context. In *WSDM*, pages 339–347, 2019.

[Dadvar *et al.*, 2013] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *ECIR*, pages 693–696, 2013.

[Dani *et al.*, 2017] Harsh Dani, Jundong Li, and Huan Liu. Sentiment informed cyberbullying detection in social media. In *ECML PKDD*, pages 52–67, 2017.

[Dinakar *et al.*, 2011] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. *The Social Mobile Web*, 11(02), 2011.

[Espelage *et al.*, 2003] Dorothy L Espelage, Melissa K Holt, and Rachael R Henkel. Examination of peer–group contextual effects on aggression during early adolescence. *Child development*, 74(1):205–220, 2003.

[Goodboy and Martin, 2015] Alan K Goodboy and Matthew M Martin. The personality profile of a cyberbully: Examining the dark triad. *Computers in human behavior*, 49:1–4, 2015.

[Gradinger *et al.*, 2011] Petra Gradinger, Dagmar Strohmeier, and Christiane Spiel. Motives for bullying others in cyberspace. *Cyberbullying in the global playground: Research from international perspectives*, 263, 2011.

[Hallac *et al.*, 2015] David Hallac, Jure Leskovec, and Stephen Boyd. Network lasso: Clustering and optimization in large graphs. In *KDD*, pages 387–396, 2015.

[Hinduja and Patchin, 2013] Sameer Hinduja and Justin W Patchin. Social influences on cyberbullying behaviors among middle and high school students. *Journal of youth and adolescence*, 42(5):711–722, 2013.

[Hosseinmardi *et al.*, 2015] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Analyzing labeled cyberbullying incidents on the instagram social network. In *SocInfo*, pages 49–66, 2015.

[Kowalski *et al.*, 2014] Robin M Kowalski, Gary W Giumetti, Amber N Schroeder, and Micah R Lattanner. Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth., 2014.

[Li *et al.*, 2014] Homa Hosseinmardi Shaosong Li, Zhili Yang, Qin Lv, Rahat Ibn Rafiq Richard Han, and Shivakant Mishra. A comparison of common users across instagram and ask. fm to better understand cyberbullying. In *BdCloud*, pages 355–362, 2014.

[Li *et al.*, 2017] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *CSUR*, 50(6):94, 2017.

[Nand *et al.*, 2016] Parma Nand, Rivindu Perera, and Abhijeet Kasture. " how bullying is this message?": A psychometric thermometer for bullying. In *COLING*, pages 695–706, 2016.

[Pennebaker *et al.*, 2001] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[Rafiq *et al.*, 2016] Rahat Ibn Rafiq, Homa Hosseinmardi, Sabrina Arredondo Mattson, Richard Han, Qin Lv, and Shivakant Mishra. Analysis and detection of labeled cyberbullying instances in vine, a video-based social network. *Social Network Analysis and Mining*, 6(1):88, 2016.

[Silva *et al.*, 2016a] Yasin N. Silva, Christopher Rich, Jaime Chon, and Lisa M. Tsosie. Bullyblocker: An app to identify cyberbullying in facebook. In *ASONAM 2016,*, pages 1401–1405, 2016.

[Silva *et al.*, 2016b] Yasin N. Silva, Christopher Rich, and Deborah Hall. Bullyblocker: Towards the identification of cyberbullying in social networking sites. In *ASONAM 2016*, pages 1377–1379, 2016.

[Smith, 2012] Peter K Smith. Cyberbullying and cyber aggression. In *Handbook of school violence and school safety*, pages 111–121. Routledge, 2012.

[Squicciarini *et al.*, 2015] A Squicciarini, S Rajtmajer, Y Liu, and Christopher Griffin. Identification and characterization of cyberbullying dynamics in an online social network. In *ASONAM*, pages 280–285, 2015.

[Wu and Huang, 2016] Fangzhao Wu and Yongfeng Huang. Personalized microblog sentiment classification via multitask learning. In *AAAI*, pages 3059–3065, 2016.

[Xu *et al.*, 2012] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *NAACL HLT*, pages 656–666, 2012.