Sentiment Informed Cyberbullying Detection in Social Media

Harsh Dani, Jundong Li(✉), and Huan Liu

Computer Science and Engineering, Arizona State University, Tempe, AZ, USA
{hdani,jundongl,huanliu}@asu.edu

Abstract. Cyberbullying is a phenomenon which negatively affects the individuals, the victims suffer from various mental issues, ranging from depression, loneliness, anxiety to low self-esteem. In parallel with the pervasive use of social media, cyberbullying is becoming more and more prevalent. Traditional mechanisms to fight against cyberbullying include the use of standards and guidelines, human moderators, and blacklists based on the profane words. However, these mechanisms fall short in social media and cannot scale well. Therefore, it is necessary to develop a principled learning framework to automatically detect cyberbullying behaviors. However, it is a challenging task due to short, noisy and unstructured content information and intentional obfuscation of the abusive words or phrases by social media users. Motivated by sociological and psychological findings on bullying behaviors and the correlation with emotions, we propose to leverage sentiment information to detect cyberbullying behaviors in social media by proposing a sentiment informed cyberbullying detection framework. Experimental results on two real-world, publicly available social media datasets show the superiority of the proposed framework. Further studies validate the effectiveness of leveraging sentiment information for cyberbullying detection.

Keywords: Cyberbullying detection · Social media
Sentiment information

1 Introduction

Cyberbullying is an increasingly important and serious social problem, which can negatively affect the individuals. It is defined as the phenomena of using the internet, cell phones and other electronic devices to willfully hurt or harass others. Due to the recent popularity and growth of social media platforms such as Facebook and Twitter, cyberbullying is becoming more and more prevalent. It has been identified as a serious national health concern by the American Psychological Association and the White House. In addition to that, according to the recent report by National Crime Prevention Council, more than 40% of teens in the US have been bullied on social media platforms [1]. The victims of cyberbullying often suffer from depression, loneliness, anxiety and low self-esteem [2]. In more tragic scenarios, the victims might attempt suicide or suffer

M. Ceci et al. (Eds.): ECML PKDD 2017, Part I, LNAI 10534, pp. 52–67, 2017.
https://doi.org/10.1007/978-3-319-71249-9_4

from interpersonal problems. Since cyberbullying is not restricted by time and place, it has more insidious effects than traditional forms of bullying [3].

Traditional mechanisms to combat cyberbullying behaviors include the development of standards and guidelines that all users must adhere to, employment of human editors to manually check the bullying behavior, the use of profane word lists, and the use of regular expressions, etc. However, these mechanisms fall short in social media since social media data is naturally dynamic [4]. As a result, the maintenance of these mechanisms is time and labor consuming. Also, they cannot scale well. Therefore, it demands the use of a principled learning framework to accurately detect new cyberbullying behaviors automatically.

The detection of cyberbullying in social media is a far more challenging task than one can expect due to the following two reasons: First, the content information in social media is short, noisy and unstructured [5]. The short and unstructured texts make traditional text representation techniques, i.e., bag-of-words very sparse and high-dimensional. As a result, traditional machine learning classifiers often cannot work well due to the curse of dimensionality [6]. Second, the users in social media intentionally obfuscate the words or phrases in the sentence to evade the manual and automatic checking. Obfuscation such as "n00b" makes it impossible for traditional mechanisms to accurately detect abusive words or phrases, leading to more false positives.

Previous psychological and sociological studies on the bullying behaviors and emotional intelligence suggest that emotional information can be used to better understand the bullying behaviors [7]. Emotional intelligence refers to the ability of an individual to accurately perceive emotion, use emotions to facilitate thought, understand and manage the emotion [8]. The lower the emotional intelligence of the user, the more likely an individual will be involved in the bullying behaviors [9]. Motivated from this insight, we investigate if the use of sentiment information of the post content could help better understand and accurately detect cyberbullying behaviors in social media.

In this paper, we attempt to perform cyberbullying detection in a supervised way by proposing a principled learning framework. More specifically, we first investigate whether sentiment information is particularly correlated with cyberbullying behaviors. Then, we discuss how to deal with short, noisy, unstructured content and how to properly leverage sentiment information for cyberbullying detection. Methodologically, we present a novel learning framework called Sentiment Informed Cyberbullying Detection (SICD). Experiments on two real-world social media datasets validate the effectiveness of the proposed framework. To summarize, we make the following contributions:

- We formally define the problem of sentiment informed cyberbullying detection in social media;
- We verify the sentiment difference between normal posts and bullying posts by comparing their sentiment score distributions;
- We present a principled learning framework which leverages sentiment information of user posts to detect cyberbullying in social media; and

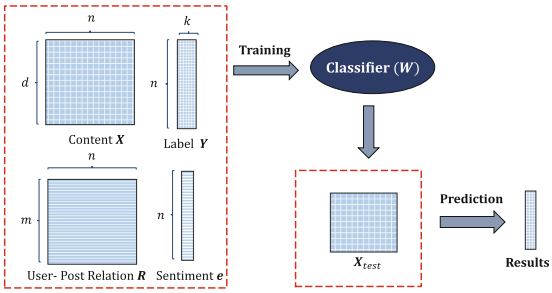


Fig. 1. Proposed sentiment informed cyberbullying detection framework.

– We perform empirical experimental studies on two real-world, publicly available social media datasets to verify the efficacy of the proposed framework.

2 Problem Definition

We first introduce the notations used in this paper. We use boldface uppercase letters (e.g., $\mathbf{A}$) to denote matrices, boldface lowercase letters to denote vectors (e.g., $\mathbf{a}$) and lowercase letters (e.g., a) to denote scalars. We denote the transpose of matrix $\mathbf{A}$ as $\mathbf{A}^T$ and the transpose of vector $\mathbf{a}$ as $\mathbf{a}^T$. $\mathrm{Tr}(\mathbf{A})$ denotes the trace of matrix $\mathbf{A}$ if it is square. The entry of matrix $\mathbf{A}$ at the row i and column j is denoted as $\mathbf{A}_{ij}$. We denote the i-th row of matrix $\mathbf{A}$ as $\mathbf{A}_{i*}$ and the j-th column as $\mathbf{A}_{*j}$. $||\mathbf{A}||_{2,1}$ denotes the $\ell_{2,1}$-norm such that $||\mathbf{A}||_{2,1} = \sum_i \sqrt{\sum_j \mathbf{A}_{ij}^2}$.

Let $\mathbf{C} = [\mathbf{X}, \mathbf{Y}]$ denote the corpus of social media posts, where $\mathbf{X} \in \mathbb{R}^{\mathrm{d}\times\mathrm{n}}$ is the content matrix of these posts, $\mathbf{Y} \in \mathbb{R}^{\mathrm{n}\times\mathrm{k}}$ is a one-hot label matrix, n is the number of posts, d is the number of features, k is the number of classes. In this work, we set $k = 2$, indicating that a post is either normal or bullying. The social media corpus $\mathbf{C}$ is generated by a set of m users, i.e., $\{u_1, u_2, \ldots, u_m\}$, $\mathbf{R} \in \mathbb{R}^{\mathrm{m}\times\mathrm{n}}$ denotes the user-post relationships (as shown in Fig. 1), $\mathbf{R}_{ij} = 1$ if post j is posted by u_i and $\mathbf{R}_{ij} = 0$ otherwise. Meanwhile, each post in $\mathbf{C}$ is associated with a sentiment score in the range of $[-1, 1]$, -1 denotes the most negative sentiment score and 1 denotes the most positive sentiment, and $\mathbf{e}$ represents the sentiment score vector for all n posts. With these notations, we now formally define the problem of sentiment informed cyberbullying detection as follows:

Given a corpus of social media posts with the content information $\mathbf{X}$ *and the label information* $\mathbf{Y}$*, the user-post relationships* $\mathbf{R}$ *and the sentiment score of posts* $\mathbf{e}$*, we aim to learn a classifier* $\mathbf{W}$ *to automatically detect whether the unseen social media posts (i.e., test data) are normal posts or bullying posts.*

3 Exploratory Data Analysis

One important motivation of the problem we study is to investigate the correlation between sentiment information and cyberbullying behaviors. We first introduce two real-world social media datasets and then present our observations from these two datasets.

3.1 Datasets

We use two publicly available social media datasets, both datasets contain labeled social media posts, i.e., the post is either labeled as normal or bullying.

Twitter is a microblogging website which allows users to post 140 characters messages. The posts in this dataset have been manually labeled as bullying or normal. This dataset has been kindly provided by Xu et al. [2].

MySpace is a social networking website which allows its registered users to view pictures, read chat and check other users' profile information. Also, each post in the dataset is manually labeled as normal or bullying. This dataset has been kindly provided by Bayzick et al. [10].

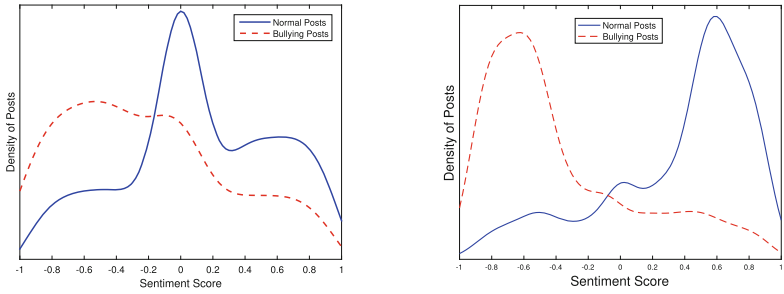
Detailed statistics of these two datasets are summarized in Table 1.

Table 1. Statistics of Twitter and MySpace datasets.

	Twitter	MySpace
# of posts	7,321	3,245
# of features	3,709	4,236
# of positive posts	2,102	950
# of negative posts	5,219	2,295
# of users	7,043	1,053
Avg. posts per user	1.04	2.98

3.2 Verifying the Sentiment Score Distribution Difference

We conduct an empirical study to verify if the sentiment distribution of the normal posts is different from the bullying posts. Particularly, we learn a distant supervision based sentiment classification model [11] on Stanford Twitter Sentiment dataset. Then we employ it to obtain the sentiment score of each post in the Twitter and MySpace datasets. The sentiment score of each post is normalized in the range of $[-1, 1]$. Figure 2(a) and (b) show the sentiment score distribution of the normal and the bullying posts in both Twitter and MySpace datasets, respectively. In these figures, X-axis shows the sentiment polarity score and Y-axis shows the density of posts. We can observe that two distributions are



(a) Twitter Dataset (b) Myspace Dataset

Fig. 2. Sentiment distribution of normal and bullying posts.

centered with different mean values. This suggests that there is a clear difference between the sentiment score distribution of the normal posts and the bullying posts, and the sentiment of bullying posts are more negative than normal posts.

3.3 Verifying Sentiment Consistency

In this subsection, we aim to investigate whether the sentiment scores of two posts with the same class labels, i.e., both posts are normal or bullying, are more similar than two randomly chosen posts. We use two-sampled t-test to verify the statistical significance of the above-stated hypothesis.

Suppose $d(p_i, p_j)$ denotes the sentiment similarity of two social media posts p_i and p_j, which can be computed by the RBF kernel. Let $\mathbf{s_c}$ and $\mathbf{s_d}$ be two vectors of the same length, each element in $\mathbf{s_c}$ denotes the sentiment similarity of two posts p_i and p_j with the same class label, and each element in $\mathbf{s_d}$ denotes the sentiment similarity of two randomly selected posts. We then use two-sampled t-test to investigate whether the sentiment similarity of two posts with the same class label is higher than two randomly chosen posts. The null hypothesis is as follows: $H_0 : \tau_c \leq \tau_d$ and the alternative hypothesis is as follows: $H_1 : \tau_c > \tau_d$, where τ_c and τ_d represent the sample means of $\mathbf{s_c}$ and $\mathbf{s_d}$, respectively.

The result of t-test, i.e., p-values obtained on Twitter and MySpace are 1.09e^{-11} and 1.028e^{-7}, respectively. It suggests that there is a strong statistical evidence (with a significance level $\alpha = 0.01$) to reject the null hypothesis. In other words, we validate the alternative hypothesis that the sentiment scores of two posts with the same class label are more similar than two randomly chosen posts. The two-sampled t-test results further pave the way to incorporate sentiment information for cyberbullying detection.

4 The Proposed Framework - SICD

In this section, we introduce the proposed SICD framework in detail. First, we present how to model the short, noisy and unstructured user post content.

Then we discuss how to model the user-post relationships and how to model sentiment information for cyberbullying detection.

4.1 Modeling Content of Social Media Posts

In order to find better text representation for cyberbullying detection, we employ unigram model with TF-IDF as feature values due to its success in cyberbullying detection [12]. Also, we perform stopwords removal and stemming.

In social media, the posts made by users are often short, noisy and unstructured. Also, these posts are not necessarily about the same topic which causes the vocabulary size to be extremely large. Hence, traditional text representation techniques such as n-grams and bag-of-words become extremely high-dimensional. Also, short text content of posts causes these feature representations to be extremely sparse. Such high-dimensional and sparse feature representations often cause poor prediction performance of traditional machine classifiers.

In recent years, sparse learning has been widely used to alleviate the negative effects of high-dimensional features to improve the prediction performance. Hence, we employ sparse learning techniques to deal with sparse, noisy and unstructured posts. More specifically, we use $\ell_{2,1}$-norm regularization term to seek a more compact feature space. The $\ell_{2,1}$-norm regularization selects a subset of relevant features across all the data instances with joint sparsity [13,14]. The classification problem then can be formulated as follows:

$$\min_{\mathbf{W}} \frac{1}{2}||\mathbf{X}^T\mathbf{W} - \mathbf{Y}||_F^2 + \lambda||\mathbf{W}||_{2,1}, \tag{1}$$

where λ is a parameter to control the feature sparsity. In the above formulation, the first term minimizes the least squared loss between post content and class labels and the second term seeks a more compact feature representation.

4.2 Modeling User-Post Relationships

Text data in social media is often linked due to the presence of various social relations, and these correlations can be explained by well-received social science theories such as Homophily [15] and Social Influence [16]. In particular, we hypothesize that if two social media posts are from the same user, they are more likely to have the same class label than two randomly chosen posts. In order to test this hypothesis, we create two equally sized vectors $\mathbf{up_c}$ and $\mathbf{up_d}$, where each element of the first vector denotes the label difference (Euclidean distance) of two posts by the same user and each element of the second vector denotes the label difference of two randomly chosen posts. We perform two-sampled t-test to investigate the above hypothesis. We form the null hypothesis as H_0: $m_c \geq m_d$ and the alternative hypothesis as H_1: $m_c < m_d$, where m_c and m_d represent the sample means of $\mathbf{up_c}$ and $\mathbf{up_d}$, respectively. The results show that there is a strong evidence (with a significance level $\alpha = 0.01$) to reject the null hypothesis.

In order to model the above mentioned user-post relationships, we propose to add a regularization term to minimize the label difference of the two posts if

they are from the same user. Specifically, we first construct an affinity matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ from matrix $\mathbf{R}$ as follows: $\mathbf{A} = \mathbf{R}^T\mathbf{R}$, such that $\mathbf{A}_{ij} = 1$ denotes that two social media posts are by the same user and $\mathbf{A}_{ij} = 0$ otherwise. With this, the user-post relationships can be modeled by minimizing the following term:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{A}_{ij}||\hat{\mathbf{Y}}_{i*} - \hat{\mathbf{Y}}_{j*}||_2^2 = \mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}_\mathbf{A}\mathbf{X}^T\mathbf{W}), \tag{2}$$

where $\hat{\mathbf{Y}} = \mathbf{X}^T\mathbf{W}$ is the predicted value of the class label $\mathbf{Y}$. $\mathbf{L}_\mathbf{A} = \mathbf{D}_\mathbf{A} - \mathbf{A}$ is the Laplacian matrix, $\mathbf{D}_\mathbf{A}$ is a diagonal matrix with $\mathbf{D}_\mathbf{A} = \sum_i \mathbf{A}_{ij}$.

4.3 Modeling Sentiment Information

Motivated by psychological and sociological findings on the correlation of emotions and bullying behaviors, we propose to incorporate sentiment information to detect cyberbullying behaviors. From Sect. 3, we have an observation that sentiment score distributions of normal posts and bullying posts are different and posts with the same label are more likely to have similar sentiment scores than two randomly chosen posts. Now we discuss how to leverage these observations to perform cyberbullying detection.

To model sentiment information of posts, we construct an undirected affinity graph $\mathbf{S} \in \mathbb{R}^{n \times n}$ where each node denotes a social media post and edge weight denotes the sentiment similarity. In this paper, we propose to construct the k-nearest neighbor graph to model the sentiment affinity between different posts. More specifically, the matrix $\mathbf{S}$ can be defined as:

$$\mathbf{S}_{ij} = \begin{cases} \exp(-\frac{||e_i - e_j||_2^2}{\sigma^2}) & \text{if } e_i \in \mathcal{N}_k(e_j) \text{ or } e_j \in \mathcal{N}_k(e_i) \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{N}_k(e_i)$ denotes the k-nearest neighbors of post p_i in terms of sentiment score. Then, we propose to model sentiment information with another Graph Laplacian [17]. The key idea is that if the sentiment scores of two posts are close to each other, their labels are similar. We formulate the above idea by minimizing the following term:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{S}_{ij}||\hat{\mathbf{Y}}_{i*} - \hat{\mathbf{Y}}_{j*}||_2^2 = \mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}_\mathbf{S}\mathbf{X}^T\mathbf{W}), \tag{3}$$

where $\mathbf{L}_\mathbf{S} = \mathbf{D}_\mathbf{S} - \mathbf{S}$ is the Laplacian matrix of the sentiment affinity matrix $\mathbf{S}$. Here, $\mathbf{D}_\mathbf{S}$ denotes the diagonal degree matrix with $\mathbf{D}_\mathbf{S} = \sum_i \mathbf{S}_{ij}$.

4.4 Sentiment Informed Cyberbullying Detection (SICD)

As illustrated from the previous sections, we employ sparse learning to model the content of the social media post. Also, we model user-post relationships and

sentiment information. By considering all the types of the information, the task of sentiment informed cyberbullying detection can be formulated as:

$$\min_{\mathbf{W}} F(\mathbf{W}) = \frac{1}{2}||\mathbf{X}^T\mathbf{W} - \mathbf{Y}||_F^2 + \lambda||\mathbf{W}||_{2,1} + \frac{\alpha}{2}\mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L_A}\mathbf{X}^T\mathbf{W}) + \frac{\beta}{2}\mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L_S}\mathbf{X}^T\mathbf{W}), \quad (4)$$

where α and β are parameters to control the contribution of user-post relationships and sentiment information, respectively. By solving the objective function in Eq. (4), we can get $\mathbf{W}$ as the learned classifier. To detect the cyberbullying behaviors on unseen social media post $\mathbf{x}$, we can use the following formulation: $\arg\max_{i\in\{bully,normal\}} \mathbf{x}^T\mathbf{W}_{*i}$.

5 Algorithmic Details

Due to the presence of the $\ell_{2,1}$-norm, the optimization problem in Eq. (4) is non-smooth but convex. Now we introduce how to solve the optimization problem along with the time complexity analysis.

5.1 Optimization Algorithm for SICD

A natural choice to solve the optimization problem in Eq. (4) is to use sub-gradient descent method [18]. However, it has a very slow convergence rate, i.e., $O(\frac{1}{\epsilon^2})$ where ϵ denotes the desired accuracy, which makes it not suitable for real-world applications. In recent years, proximal gradient descent [19,20] has been widely used to solve large-scale non-smooth convex optimization problems, where the objective function can be separated into both smooth and non-smooth parts. In our scenario, $||\mathbf{W}||_{2,1}$ is the non-smooth part and the other terms form the smooth part $f(\mathbf{W})$. In each iteration of proximal gradient descent, $F(\mathbf{W})$ is linearized around the current estimate $\mathbf{W}_t$, where t indicates the t-th iteration. In particular, $\mathbf{W}$ is updated by solving the following optimization problem:

$$\mathbf{W}_{t+1} = \arg\min_{\mathbf{W}} \mathcal{G}_{\eta_t}(\mathbf{W}, \mathbf{W}_t). \quad (5)$$

$\mathcal{G}_{\eta_t}(\mathbf{W}, \mathbf{W}_t)$ is defined as:

$$\mathcal{G}_{\eta_t}(\mathbf{W}, \mathbf{W}_t) = f(\mathbf{W}_t) + \langle \nabla f(\mathbf{W}_t), \mathbf{W} - \mathbf{W}_t \rangle + \frac{\eta_t}{2}||\mathbf{W} - \mathbf{W}_t||_F^2 + \lambda||\mathbf{W}||_{2,1} \quad (6)$$

where η_t is the step size that can be determined by the backtracking line search algorithm. $\langle \mathbf{A}, \mathbf{B} \rangle$ denotes the dot product between two matrices $\mathbf{A}$ and $\mathbf{B}$: $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}^T\mathbf{B})$. The gradient of the smooth part $f(\mathbf{W})$ is formulated as:

$$\nabla f(\mathbf{W}_t) = \mathbf{X}\mathbf{X}^T\mathbf{W}_t - \mathbf{X}\mathbf{Y} + \alpha\mathbf{X}\mathbf{L_A}\mathbf{X}^T\mathbf{W}_t + \beta\mathbf{X}\mathbf{L_S}\mathbf{X}^T\mathbf{W}_t. \quad (7)$$

In Eq. (6), we ignore the terms that are not related to $\mathbf{W}$ and the objective function boils down to the following optimization problem:

$$\mathbf{W}_{t+1} = \pi_{\eta_t}(\mathbf{W}_t) = \arg\min_{\mathbf{W}} \frac{1}{2}||\mathbf{W} - \mathbf{U}_t||_F^2 + \frac{\lambda}{\eta_t}||\mathbf{W}||_{2,1}, \tag{8}$$

where $\mathbf{U}_t = \mathbf{W}_t - \frac{1}{\eta_t}\nabla f(\mathbf{W}_t)$. The above problem can be further decomposed into k sub-problems. Each sub-problem can be formally formulated as follows:

$$\mathbf{w}_{t+1}^i = \arg\min_{\mathbf{w}^i} ||\mathbf{w}^i - \mathbf{u}_t^i||_2^2 + \frac{\lambda}{\eta_t}||\mathbf{w}^i||_2, \tag{9}$$

where the $\mathbf{w}_{t+1}^i$, $\mathbf{w}^i$ and $\mathbf{u}_t^i$ are the i-th row of the matrix $\mathbf{W}_{t+1}$, $\mathbf{W}$ and $\mathbf{U}_t$, respectively. Given the value of λ, the Euclidean projection of the above optimization problem has a closed-form solution, which can be formulated as:

$$\mathbf{w}_{t+1}^i = \begin{cases} (1 - \frac{\lambda}{\eta_t||\mathbf{u}_t^j||_2})\mathbf{u}_t^j; & \text{if } ||\mathbf{u}_t^j||_2 \geq \frac{\lambda}{\eta_t}, \\ \mathbf{0}; & \text{otherwise.} \end{cases} \tag{10}$$

Since the algorithm described above has closed-form Euclidean projection [20], hence it has the same convergence rate (i.e., $\frac{1}{\epsilon}$) as traditional gradient descent algorithms for smooth convex optimization problems. As discussed in [20], the proximal algorithm can be further accelerated to achieve the optimal convergence rate of $O(\frac{1}{\sqrt{\epsilon}})$ by employing Nestrov's method [21].

5.2 Time Complexity Analysis

Given a corpus of $\mathbf{C}$ with n social media posts and a feature dimension of d, it requires $O(nd)$ operations to obtain the gradient of the least squared formulation. The Euclidean projection for the $\ell_{2,1}$-norm according to Eq. (10) requires $O(2n)$ operations [20]. Third, the Laplacian regularization for the modeling of user-post relationships requires $O(nd)$. Similarly, the Laplacian regularization for the modeling of sentiment information also requires $O(nd)$. Also, by employing the Nestrov's accelerated method, we can achieve the optimal convergence rate of $O(\frac{1}{\sqrt{\epsilon}})$. Hence, the total time complexity of the proposed Algorithm is $O(\frac{1}{\sqrt{\epsilon}}(nd + 2n + nd + nd)) = O(\frac{1}{\sqrt{\epsilon}}(nd))$.

6 Experiments

In this section, we perform experiments to evaluate the effectiveness of the proposed SICD framework. After introducing the experimental settings, we present the detailed experimental results.

6.1 Experimental Settings

We follow standard experimental settings [12] to evaluate the performance of the proposed SICD framework. To avoid the bias brought by imbalanced class distributions, we use AUC and F1-measure as the classification metrics.

There are three positive parameters involved in our framework. λ controls the contribution of the sparse regularization. α controls the contribution of user-post relationships and β controls the contribution of sentiment information modeling. In the experiments, we set these parameters as $\lambda = 0.1$, $\alpha = 0.1$, $\beta = 0.05$, and $k = 20$ for the k-nearest neighbor in Eq. (3) by using grid search strategies.

6.2 Performance Evaluation

We compare our proposed SICD framework with the following baseline methods:

- **LS:** Traditional linear classification method with least squared loss [22].
- **Lasso:** This is a supervised sparse learning method [22] which uses ℓ_1-norm sparse regularization on the basis of least squared loss.
- **MF:** We perform NMF [23] on the content information for a compact representation and then apply SVM.
- **POS:** This method uses TF-IDF features, POS-tags of the bigrams, and the list of profane words as feature sets and then classifies posts using SVM [24].
- **USER:** This method uses TF-IDF features, and user related features such as gender and age as feature sets and then classifies posts using SVM [12].

For the *USER* baseline, if the user did not provide age or gender information, we impute the age information by the mean value and gender information by the most frequent value of others. For all methods, we perform five-fold cross-validation and report the average results. Particularly, we first divide 80% of the data as training data and the remaining 20% as the test set. The Tables 2 and 3 summarize the results on the Twitter dataset and MySpace dataset, respectively. It should be noted that in these tables, the training ratio is varied among the 80% training data.

We draw the following observations from these two tables:

- *SICD* consistently outperforms other baseline methods on both datasets with the varied training ratio. We also perform pairwise Wilcoxon signed-rank test [25] between *SICD* and other baselines, the results show that SICD is statistically significant better (with significance level $\alpha = 0.01$).
- *Lasso* and *MF* both achieves better performance than *LS*. It indicates that performing dimensionality reduction on the original content matrix can reduce the noisy information contained and helps improve the performance.
- As we increase the training ratio from 10% to 90%, the performance of *SICD* tends to increase gradually. It shows that more training data helps achieve better performance on the cyberbullying detection problem.
- *POS* outperforms *LS*, *Lasso* and *MF*, which indicates that POS tags of the frequent bi-grams and the list of profane words are often good indicators of cyberbullying behaviors.

Table 2. Classification evaluation of different methods on Twitter data.

	Train ratio	10%	20%	30%	40%	50%	60%	70%	80%	90%
F1	LS	0.4057	0.4105	0.4128	0.4264	0.4454	0.4519	0.4586	0.4662	0.4724
	Lasso	0.4635	0.5254	0.5734	0.5783	0.5870	0.5927	0.6039	0.6120	0.6187
	MF	0.5090	0.5197	0.5785	0.5819	0.5882	0.5916	0.5974	0.6008	0.6225
	POS	0.4934	0.5279	0.5812	0.5864	0.5985	0.6023	0.6104	0.6191	0.6247
	USER	0.4789	0.5190	0.5797	0.5805	0.5820	0.5939	0.6089	0.6178	0.6235
	SICD	**0.5601**	**0.5965**	**0.6127**	**0.6265**	**0.6354**	**0.6445**	**0.6697**	**0.6894**	**0.7056**
AUC	LS	0.6103	0.6142	0.6168	0.6259	0.6309	0.6338	0.6392	0.6435	0.6519
	Lasso	0.6419	0.6934	0.7219	0.7234	0.7297	0.7318	0.7532	0.7617	0.7698
	MF	0.6567	0.6745	0.7261	0.7281	0.7309	0.7335	0.7368	0.7397	0.7446
	POS	0.6497	0.6915	0.7245	0.7310	0.7391	0.7426	0.7469	0.7583	0.7624
	USER	0.6389	0.6867	0.7236	0.7295	0.7338	0.7431	0.7446	0.7516	0.7603
	SICD	**0.7049**	**0.7369**	**0.7567**	**0.7684**	**0.7869**	**0.7934**	**0.7977**	**0.8051**	**0.8169**

Table 3. Classification evaluation of different methods on MySpace data.

	Train ratio	10%	20%	30%	40%	50%	60%	70%	80%	90%
F1	LS	0.3937	0.4112	0.4370	0.4432	0.4691	0.4807	0.4916	0.5018	0.5075
	Lasso	0.3960	0.4338	0.4625	0.4759	0.4811	0.4925	0.5046	0.5120	0.5235
	MF	0.4145	0.4427	0.4861	0.4917	0.5109	0.5164	0.5254	0.5364	0.5446
	POS	0.4390	0.4502	0.4745	0.4818	0.4897	0.5012	0.5191	0.5286	0.5341
	USER	0.4267	0.4478	0.4723	0.4789	0.4842	0.4981	0.5085	0.5256	0.5320
	SICD	**0.4928**	**0.5086**	**0.5301**	**0.5572**	**0.5691**	**0.5791**	**0.5886**	**0.6071**	**0.6105**
AUC	LS	0.6096	0.6138	0.6236	0.6284	0.6348	0.6408	0.6485	0.6516	0.6547
	Lasso	0.6106	0.6219	0.6331	0.6378	0.6412	0.6501	0.6521	0.6587	0.6625
	MF	0.6147	0.6276	0.6487	0.6495	0.6569	0.6573	0.6639	0.6748	0.6837
	POS	0.6248	0.6314	0.6385	0.6418	0.6479	0.6509	0.6576	0.6657	0.6517
	USER	0.6210	0.6303	0.6349	0.6392	0.6441	0.6487	0.6461	0.6625	0.6658
	SICD	**0.6509**	**0.6549**	**0.6681**	**0.6915**	**0.6920**	**0.7224**	**0.7296**	**0.7404**	**0.7539**

– Similarly, *USER* outperforms *LS*, *Lasso* and *MF* which indicates that adding user based features such as gender and age helps improve the classification performance. However, it is inferior to the proposed *SICD*, one potential reason is that the age or gender information is often scarcely available in social media due to privacy reasons [26,27].
– Figure 3 demonstrates the prediction by *SICD* visually for the Twitter dataset. The top words for the bullying posts according to the ground truth and prediction by *SICD* are described in the top-left and bottom-left part of Fig. 3. Similarly, the top words for the normal post according to the ground truth and SICD are described in the top-right and bottom-right part of the Fig. 3. As we can observe, there is a significant overlap of the words in ground truth and prediction by *SICD* which visually demonstrates the effectiveness of the proposed framework.

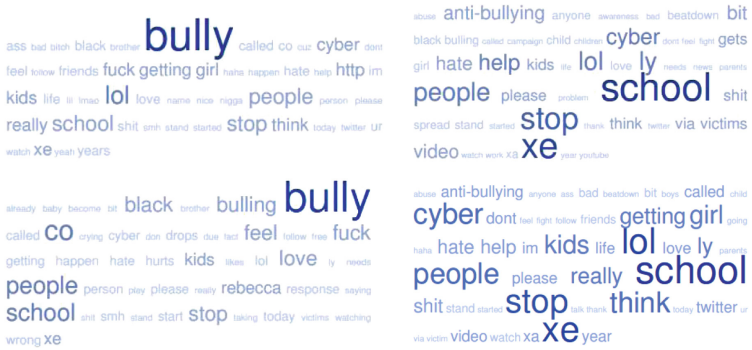


Fig. 3. Prominent words for Twitter dataset.

6.3 Impact of Sentiment Information

In order to investigate the impact of sentiment information on cyberbullying detection, we assess the effectiveness of different types of information in SICD. In particular, we compare our proposed method with the following methods:

Table 4. Impact of sentiment information in Twitter and MySpace datasets.

Methods	Twitter		MySpace	
	F1	AUC	F1	AUC
Sentiment	0.2014	0.5214	0.2106	0.5197
Content	0.4724	0.6519	0.5075	0.6547
Content+UPR	0.6544	0.7921	0.5908	0.7032
Content+SENT	0.6298	0.7846	0.5894	0.6891
SICD	**0.7056**	**0.8169**	**0.6105**	**0.7539**

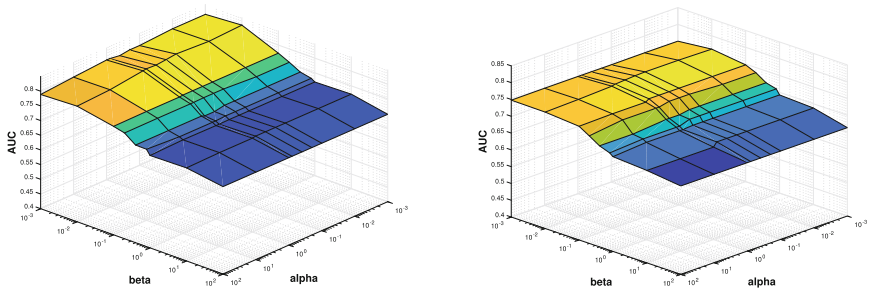
- **Content:** This is the traditional Least Squared classification model where only content information $\mathbf{X}$ is used.
- **Sentiment:** We first compute the sentiment score of the each post and then calculate its distance from the mean of the bullying and normal posts groups. The post is classified into the group with the shorter distance.
- **Content+UPR:** This method is a variant of our method, where the sentiment regularization is removed.
- **Content+SENT:** This method is a variant of our method, where the user-post relationship regularization is removed.

The experimental results on Twitter and MySpace are in Table 4. We have the following observations:

- With all the types of the information considered, *SICD* achieves the best cyberbullying detection performance.
- The *Sentiment* method achieves the worst performance. It indicates that although we observe the difference in the sentiment scores of the normal posts and the bullying posts, we cannot just use this information to detect cyberbullying. *Content* achieves better performance compared to *Sentiment*. It suggests that content information is the most effective source of information to perform cyberbullying detection.
- The *Content+UPR* and the *Content+SENT* achieves better performance than *Content* and *Sentiment*. It indicates that integration of either user-post relationships or sentiment information helps achieve better performance compared to traditional text-based cyberbullying detection methods.

6.4 Parameter Sensitivity

Our proposed SICD has two important parameters: α and β. The parameter α and β control the contribution of user-post relationships and sentiment information, respectively. To better understand the effects of these two parameters, we vary the values of α and β as $\{0.001, 0.01, 0.1, 0.2, 0.4, 0.8, 1.0, 10.0, 100.0\}$ and report the classification performance on both datasets. The classification results w.r.t. AUC are shown in Fig. 4(a) and (b).



(a) Twitter Dataset (b) Myspace Dataset

Fig. 4. Impact of the parameters α and β on the proposed framework

It can be observed that when β is around 0.01, SICD achieves the best performance. When we gradually increase the value of β, the cyberbullying detection performance first increases and then keeps stable. We can also observe that SICD is not very sensitive to the parameter α, thus we can tune it in a safe range. In usual practice, the parameters α and β should be in the range of $[0.001, 1]$.

7 Related Work

In this section, we briefly introduce the related work in cyberbullying detection and sentiment analysis in social media.

We first briefly review the related literature of detecting cyberbullying behaviors in social media. Dinakar et al. [24] proposed the problem of modeling textual information to detect cyberbullying behaviors on the web. They used concatenation of several feature sets, such as TF-IDF features, POS tags of frequent bigrams and list of profane words to predict the presence of bullying. Dadvar et al. [12] used the user related features such as gender, age to show that such user based features can be used to improve the prediction performance. Xu et al. [2] proposed several models such as BoW based models, LSA (Latent Semantic Analysis) based and LDA based models to predict the bullying behaviors in social media. However, most of them presented an exploratory study rather than providing a principled learning framework. Dinakar et al. [1] presented a common sense based reasoning approach to construct the bullying knowledge base and incorporated it into the cyberbullying detection framework. However, the construction of such knowledge base for each dataset is a labor intensive work. Also, real-world social networks often evolve over time which makes the development of this knowledge base even more difficult and time-consuming. In the later work, Squicciarini et al. [3] presented an approach based on pairwise interactions between users in social networks to identify the bullying users. Particularly, the authors considered interactions of the cyberbullies with normal users in addition to the bag-of-words text analysis.

Another research area related to our work is sentiment analysis in social media. Traditional sentiment analysis has been extensively studied in literature. It has been applied to different corpus such as product reviews [28–30], movie reviews [31,32] and newspaper articles [33]. Recently, the sentiment analysis in social media has received increasing attention since social media is an opinion-rich resource. Sentiment analysis finds many applications in social media realm [34–36] such as poll-rating prediction [37], event detection and prediction [38]. However, the use of sentiment analysis to detect malicious behaviors in social media is limited. One particular use of sentiment analysis to detect malicious posts from social media is done by Cambria et al. [39]. [40] used sentiment analysis to identify various emotions from the bullying behaviors. More specifically, the authors used a trained model and applied it to the Twitter dataset to discover various emotional patterns. However, this work is different from ours as we leverage the sentiment score difference between normal posts and bullying posts and proposed a principle learning framework.

8 Conclusion and Future Work

In this paper, we study the problem of sentiment informed cyberbullying detection in social media. The unique characteristics of the social media data and intentional obfuscation of the abusive words present unique challenges for cyberbullying detection. Motivated by the psychological and sociological findings, we propose to leverage sentiment information to help detect cyberbullying behaviors in social media. First, we conduct an exploratory data analysis on the

Twitter and MySpace datasets and observe that sentiment information can be potentially useful for cyberbullying detection. Methodologically, we propose a principled sparse learning framework by incorporating sentiment information and user-post relationships. Finally, we conduct extensive experiments on two real-world datasets. The experimental results show the effectiveness of the proposed model as well as the impact of sentiment information.

There are many future directions. Most of the work done so far in cyberbullying detection has been found in the English language. However, it is important to develop methods to handle other languages as well. Another future work is to investigate the impact of the sarcasm information hidden in the posts for cyberbullying detection.

Acknowledgements. This material is based upon work supported by, or in part by, the NSF grant 1614576, and the ONR grant N00014-16-1-2257.

References

1. Dinakar, K., Jones, B., Havasi, C., Lieberman, H., Picard, R.: Common sense reasoning for detection, prevention, and mitigation of cyberbullying. TiiS **2**(3), 18 (2012)
2. Xu, J.M., Jun, K.S., Zhu, X., Bellmore, A.: Learning from bullying traces in social media. In: NAACL-HLT, pp. 656–666 (2012)
3. Squicciarini, A., Rajtmajer, S., Liu, Y., Griffin, C.: Identification and characterization of cyberbullying dynamics in an online social network. In: ASONAM, pp. 280–285 (2015)
4. Li, J., Hu, X., Jian, L., Liu, H.: Toward time-evolving feature selection on dynamic networks. In: ICDM, pp. 1003–1008 (2016)
5. Baldwin, T., Cook, P., Lui, M., MacKinlay, A., Wang, L.: How noisy social media text, how diffrnt social media sources? In: IJCNLP, pp. 356–364 (2013)
6. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H.: Feature selection: a data perspective. arXiv preprint arXiv:1601.07996 (2016)
7. Kokkinos, C.M., Kipritsi, E.: The relationship between bullying, victimization, trait emotional intelligence, self-efficacy and empathy among preadolescents. Soc. Psychol. Educ. **15**(1), 41–58 (2012)
8. Mayer, J.D., Roberts, R.D., Barsade, S.G.: Human abilities: emotional intelligence. Annu. Rev. Psychol. **59**, 507–536 (2008)
9. Mckenna, J., Webb, J.A.: Emotional intelligence. Br. J. Occup. Ther. **76**(12), 560–561 (2013)
10. Bayzick, J., Kontostathis, A., Edwards, L.: Detecting the presence of cyberbullying using computer software. In: WebSci (2011)
11. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, vol. 1, p. 12 (2009)
12. Dadvar, M., De Jong, F.: Cyberbullying detection: a step toward a safer internet yard. In: WWW, pp. 121–126 (2012)
13. Li, J., Wu, L., Zaïane, O.R., Liu, H.: Toward personalized relational learning. In: SDM, pp. 444–452 (2017)
14. Li, J., Dani, H., Hu, X., Liu, H.: Radar: residual analysis for anomaly detection in attributed networks. In: IJCAI (2017)

15. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. Annu. Rev. Sociol. **27**, 415–444 (2001)
16. Marsden, P.V., Friedkin, N.E.: Network studies of social influence. Sociol. Methods Res. **22**(1), 127–151 (1993)
17. Chung, F.R.: Spectral Graph Theory, vol. 92. American Mathematical Society, Providence (1997)
18. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
19. Ji, S., Ye, J.: An accelerated gradient method for trace norm minimization. In: ICML, pp. 457–464 (2009)
20. Liu, J., Ji, S., Ye, J.: Multi-task feature learning via efficient l2, 1-norm minimization. In: UAI, pp. 339–348 (2009)
21. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer, New York (2013). https://doi.org/10.1007/978-1-4419-8853-9
22. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning
23. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: NIPS, pp. 556–562 (2001)
24. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. In: The Social Mobile Web, pp. 11–17 (2011)
25. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. JMLR **7**, 1–30 (2006)
26. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on Twitter. In: EMNLP, pp. 1301–1309 (2011)
27. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: WSDM, pp. 251–260 (2010)
28. Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: WSDM, pp. 231–240 (2008)
29. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: KDD, pp. 168–177 (2004)
30. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
31. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: ACL, p. 271 (2004)
32. Zhuang, L., Jing, F., Zhu, X.Y.: Movie review mining and summarization. In: CIKM, pp. 43–50 (2006)
33. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: EMNLP, pp. 79–86 (2002)
34. Hu, X., Tang, L., Tang, J., Liu, H.: Exploiting social relations for sentiment analysis in microblogging. In: WSDM, pp. 537–546 (2013)
35. Hu, X., Tang, J., Gao, H., Liu, H.: Unsupervised sentiment analysis with emotional signals. In: WWW, pp. 607–618 (2013)
36. Cheng, K., Li, J., Tang, J., Liu, H.: Unsupervised sentiment analysis with signed social networks. In: AAAI, pp. 3429–3435 (2017)
37. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: linking text sentiment to public opinion time series. In: ICWSM, pp. 1–2 (2010)
38. Bollen, J., Mao, H., Pepe, A.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: ICWSM, pp. 450–453 (2011)
39. Cambria, E., Chandra, P., Sharma, A., Hussain, A.: Do not feel the trolls. ISWC (2010)
40. Xu, J.M., Zhu, X., Bellmore, A.: Fast learning for sentiment analysis on bullying. In: Workshop on Issues of Sentiment Discovery and Opinion Mining (2012)