Sentiment Informed Cyberbullying Detection in Social Media

by

Harsh Dani

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved January 2017 by the
Graduate Supervisory Committee:

Huan Liu, Chair
Hanghang Tong
Jingrui He

ARIZONA STATE UNIVERSITY

May 2017

ABSTRACT

Cyberbullying is a phenomenon which negatively affects individuals. Victims of the cyberbullying suffer from a range of mental issues, ranging from depression to low self-esteem. Due to the advent of the social media platforms, cyberbullying is becoming more and more prevalent. Traditional mechanisms to fight against cyberbullying include use of standards and guidelines, human moderators, use of blacklists based on profane words, and regular expressions to manually detect cyberbullying. However, these mechanisms fall short in social media and do not scale well. Users in social media use intentional evasive expressions like, obfuscation of abusive words, which necessitates the development of a sophisticated learning framework to automatically detect new cyberbullying behaviors. Cyberbullying detection in social media is a challenging task due to short, noisy and unstructured content and intentional obfuscation of the abusive words or phrases by social media users. Motivated by sociological and psychological findings on bullying behavior and its correlation with emotions, we propose to leverage the sentiment information to accurately detect cyberbullying behavior in social media by proposing an effective optimization framework. Experimental results on two real-world social media datasets show the superiority of the proposed framework. Further studies validate the effectiveness of leveraging sentiment information for cyberbullying detection.

To my Parents, Grand Parents and Aastha

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Huan Liu for his immense support and guidance throughout my masters degree here at ASU. His pragmatism and advice helped me through out my journey here. I am sure that lessons learnt will help me in my future endeavors.

I would also like to thank my committee members Dr. Hanghang Tong and Dr. Jingrui He for being on my committee without hesitation. Suggestions and critical comments from them helped to improve this significantly.

I would also like to thank Jundong Li, my friend and collaborator, for instilling the value of hard work in me and answering all my queries patiently throughout this thesis work. I would also like to thank Fred Morstatter for all the encouragement and helping me write my first paper. I would like thank Isaac Jones for helping me proofread the thesis. I also thank DMML lab members Tahora H. Nazer, Kian Fakhri, Liang Wu, Xia Hu, Rob Trevino, Kai Shu, Suhas Ranganath, Justin Sampson, Kewei Cheng, Ghazaleh Beigi and Suhang Wang. I learnt a lot and made great memories.

Life in Tempe would have been not so great if I didn't have an amazing group of friends. Thanks Anuj, Rushi, Maunil, Prashil, Kathan, Akash, Bhavin, and Dhruvil - couldn't have done it without you guys.

Last but not the least, I would like to thank my mom, dad, grandmother, grandfather and Aastha for encouraging me and standing throughout my life - words don't do justice to everything you've done for me.

TABLE OF CONTENTS

	Page
LIST OF TABLES	vi
LIST OF FIGURES	vii
CHAPTER	
1 INTRODUCTION	1
2 RELATED WORK	4
2.1 Bullying Detection on the Web	4
2.2 Sentiment Classification in Social Media	8
2.3 Abusive Language Detection in Social Media	9
3 PROBLEM DEFINITION	12
4 EXPLORATORY DATA ANALYSIS	14
4.1 Datasets	14
4.2 Verifying Sentiment Score Distribution	15
4.3 Verifying Sentiment Consistency	16
5 PROPOSED FRAMEWORK - SICD	18
5.1 Modeling Content of Social Media Posts	18
5.2 Modeling User-Post Relationship	19
5.3 Modeling Sentiment Information	20
5.4 Sentiment Informed Cyberbullying Detection (SICD)	21
6 ALGORITHMIC DETAILS	23
6.1 Optimization Algorithm for SICD	23
6.2 Accelerated Proximal Algorithm for SICD	25
6.3 Convergence Analysis	26
6.4 Time-Complexity Analysis	28
7 EXPERIMENTS	29

CHAPTER	Page
7.1 Experimental Settings.....	29
7.2 Performance Evaluation	30
7.3 Impact of Sentiment Information.....	34
7.4 Parameter Sensitivity	36
8 DISCUSSIONS AND FUTURE WORK	38
REFERENCES	40

LIST OF TABLES

Table	Page
2.1 Overview of Related Work	5
4.1 Statistics of Twitter and MySpace Datasets	15
7.1 The F1-score and AUC score of Different Methods on Twitter Data with Varying Training Data Sizes.	30
7.2 The F1-score and AUC score of Different Methods on MySpace Data with Varying Training Data Sizes.	31
7.3 Impact of Sentiment Information in Twitter and MySpace Datasets....	35

LIST OF FIGURES

Figure	Page
3.1 Overview of the Proposed Framework	12
4.1 Sentiment Score Distribution of Normal Posts and Bullying Posts in the Twitter Dataset.	16
7.1 Prominent Words for Twitter Dataset	32
7.2 Impact of the User-Post Relationship (α) and Sentiment (β) on the Proposed Framework in Twitter Dataset.....	37

Chapter 1

INTRODUCTION

Cyberbullying is an increasingly important and serious social problem, which can negatively affect 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 ¹ and the White House ². In addition to that, according to the recent report by National Crime Prevention Council, more than 40% of the teens in the US have been bullied on various social media platforms Dinakar *et al.* (2012). The victims of the cyberbullying often suffer from depression, loneliness, anxiety, and low self-esteem Xu *et al.* (2012a). In more tragic scenarios, the victims might attempt suicide or suffer from interpersonal problems. Since cyberbullying is not restricted by time and place, it has more insidious effects than traditional forms of bullying Squicciarini *et al.* (2015).

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 for bullying behavior, the use of profane word lists, and the use of regular expressions. However, these mechanisms fall short in social media. As a result, the maintenance of these mechanisms is time and labor consuming. Also, they cannot scale well. Therefore, it necessitates the use of a learning framework to accurately detect new cyberbullying instances automatically.

¹The American Psychological Association Resolution on Bullying Among Children and Youth

²The White House Conference on Bullying Prevention

The detection of the cyberbullying in social media is a far more challenging task than expected for two reasons: First, the content information in social media is short, noisy and unstructured Baldwin *et al.* (2013). This short and unstructured text make traditional text representation techniques like, bag-of-words to be very sparse and high-dimensional. This sparse and high-dimensional feature space cause poor prediction accuracy of the machine learning classifiers. Hu *et al.* (2013b) 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 difficult for traditional mechanisms to accurately detect abusive words or phrases. Hence, these mechanisms can lead to more false positives.

Previous psychological and sociological studies on bullying behavior and emotional intelligence suggest that emotional information can be used to better understand bullying behavior Kokkinos and Kipritsi (2012). Emotional intelligence refers to the ability of an individual to accurately perceive emotion, use emotions to facilitate thought, and understand and manage emotion Mayer *et al.* (2008). The lower the emotional intelligence of the user, the more likely an individual will be involved in the bullying behavior Mckenna and Webb (2013). 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 thesis, we attempt to perform cyberbullying detection in a supervised way by proposing a learning framework. More specifically, we first investigate whether sentiment information is correlated with cyberbullying behavior. Then, we discuss how to deal with short, noisy, unstructured content and how to properly leverage sentiment information for cyberbullying detection. Afterwards, we present a novel optimization framework called Sentiment Informed Cyberbullying Detection (SICD). Extensive experiments on two real-world social media datasets validate the effectiveness of the

proposed framework. To the best of our knowledge, this is the first attempt to leverage sentiment information to detect cyberbullying behaviors with a learning model. To summarize, we make the following contributions:

- Formally define the problem of sentiment informed cyberbullying detection in social media;
- Verify that there exists a difference of sentiment between normal posts and bullying posts by comparing the sentiment score distribution;
- Present a novel framework that leverages sentiment information of the post to detect cyberbullying in social media; and
- Perform extensive experimental studies on two real-world, publicly available social media datasets, namely Twitter and MySpace, to verify the efficacy of the proposed framework.

The remainder of the thesis is organized as follows: In chapter 2, we briefly introduce the related work. In chapter 3, we introduce the notations and formally define the problem of cyberbullying detection. In chapter 4, we perform exploratory data analysis to investigate the impact of sentiment information for the cyberbullying detection. In chapter 5, we propose a novel framework to detect cyberbullying behavior in social media by exploiting sentiment information of the post content. In chapter 6, we present the optimization solutions for the proposed cyberbullying detection framework. In chapter 7, we present experiments on two real-world social media datasets to evaluate the proposed framework. Finally, in chapter 8, we conclude the thesis and give directions for future work.

Chapter 2

RELATED WORK

In this thesis, we consider a novel problem of sentiment informed cyberbullying detection in social media. There are several research directions which are directly related to our work.

2.1 Bullying Detection on the Web

We first present related literature on detecting cyberbullying from psychological and sociological perspectives and then later we introduce computational methods to detect cyberbullying on the web and other social media platforms. The standard approach in psychological and sociological science is to conduct personal surveys. Most of the research conducted by psychologists or social scientists is in the school environment via self, peer, and teacher surveys about the experiences of individuals as victim and perpetrator Card and Hodges (2008). However, such manual surveys are usually conducted only once and the sample size is usually in hundreds. The participants usually write 3 to 4 lines about their experiences Nishina and Bellmore (2010). However, some researchers have tried more longitudinal studies that span several months to years Nishina and Juvonen (2005); Nylund *et al.* (2007). Despite surveys being the most widely adopted method, there are several limitations of this method. First, the sample size covered by such surveys is tiny compared to the whole population. If the study is conducted in a specific geographical region the conclusion of the survey might be biased towards that geographical region. Second, the schools

Table 2.1: Overview of Related Work

Authors and Year	Overview of Methodology
Dinakar <i>et al.</i> (2011)	TF-IDF features, POS tags of the frequent bigrams, list of profane words as features classified using SVM.
Dadvar and De Jong (2012)	TF-IDF features with gender features, and user’s contextual features.
Xu <i>et al.</i> (2012a)	Bag-of-words, LSA, and LDA based models to predict bullying traces in social media.
Dinakar <i>et al.</i> (2012)	TF-IDF features, with common stereotypical words, word list based on negative words, and open mind common sense knowledge base
Squicciarini <i>et al.</i> (2015)	Interactions of cyberbullies with normal user in addition to social network features, content features

based experiences of bullying are recorded, which places more emphasis on particular age-groups while neglecting other age-groups Sui (2015).

Computational study of the cyberbullying behavior largely remains unexplored. However, there are some exceptions Dadvar and De Jong (2012); Dadvar *et al.* (2013); Dinakar *et al.* (2011); Lieberman *et al.* (2011); Xu *et al.* (2012a). All of this work looks at language patterns and find the messages that are most likely to exhibit the cyberbullying behavior. More specifically, Dinakar *et al.* (2011) proposed the problem of modeling textual information to detect cyberbullying instances on the web. Using the PHP API of YouTube, authors downloaded about 50,000 video comments involving sensitive topics like race and culture, sexuality, and intelligence. The comments were further clustered into categories mentioned above and was later converted into three

different datasets. 1500 comments from each of the clusters were annotated manually. From each of the datasets, authors removed stopwords, performed stemming and unimportant characters and sequences. The authors used various features such as TF-IDF, POS tags of the frequent bigrams, lists of profane words Ortony *et al.* (1987) and label specific features that were commonly observed in the training data. The authors then used various classifiers to learn a model including naive bayes, rule-based Jrip, J4.8 decision tree, and SVM. The authors observed that SVM outperforms other classifiers in terms of prediction accuracy. Another observation authors made was that the most discriminative features were the label-specific features. However, these label specific features are usually specific to dataset. The task of building label specific features can be cumbersome for large datasets. Also, due to the sparse nature of social media data it becomes difficult to obtain label-specific unigrams and bigrams.

Dadvar and De Jong (2012) considers the problem of predicting presence of bullying in social media posts. The authors hypothesize that adding gender specific features and user’s contextual features will help improve the prediction performance of the classifier. To validate their hypothesis, authors obtained data from MySpace which is comprised of more than 381,000 comments from predefined topics. The authors employ gender specific features and construct two datasets, one for male and another for female. They also use unigram features with TF-IDF weights with gender based features. They train a SVM classifier on the proposed feature set and observed that adding user and gender specific features improved the prediction performance.

In the later work, Dadvar *et al.* (2013) added contextual features of user such as previous posts of the users, their interaction, use of profane words in the previous posts to improve the performance of the classifier. They used comments from YouTube as their data. The authors observed that content based features alone are not sufficient to detect cyberbullying instances with high accuracy. However, the addition of the user-

based features and number of profanities in user’s comment history further improved the classification performance.

Xu *et al.* (2012a) proposed several models such as Bag of Words based, Latent Semantic Analysis based, and LDA based models to predict bullying traces in social media. The main contribution of the paper is to show that, with appropriate natural language processing techniques, social media can be a very rich source to study bullying in both physical and cyber worlds. The authors formulate problem as several natural language processing task such as text categorization, role labeling, and latent topic modeling. In this work, authors aim to present an exploratory study to detect the presence of bullying traces rather than providing a principled learning framework to classify each individual post.

Dinakar and colleagues Dinakar *et al.* (2012) presents a common sense based reasoning approach to construct a bullying specific knowledge base and incorporated it into natural language based cyberbullying detection framework. Authors use five primary feature set. The general features include TF-IDF weighted unigram features, frequently occurring stereotypical words. The second feature set include feature include lexicons that exhibit negative effects. The third feature-sets include the Part-Of-Speech (POS) tags of the most frequent bigrams observed in the training dataset. The authors build label-specific binary features by observing language patterns. Additionally, authors used the Open Mind Common Sense Singh *et al.* (2002) knowledge base to provide intuition to the classifier about language patterns. The OMCS knowledge base helps classifiers build relations between real-world entities. However, real world social networks evolve over time. New content patterns emerge and the old ones fade away Li *et al.* (2015) which makes the development and maintenance of such knowledge bases even more difficult and time-consuming, particularly in social media.

In the later work, Squicciarini *et al.* (2015) presented an approach which charac-

terizes the influence/influenced relationship in the bullying behavior. In this work, authors rely on content features, user’s contextual features, and social network features to model interaction between cyberbullies and normal users. The authors use two datasets, namely MySpace and Formspring, to validate their hypothesis. The authors perform experiments using the proposed features and observe that content features are important to identify cyberbullies and cyberbullying dynamics but user features are important to detect cyberbullies. However, user features do not really help to identify cyberbullying dynamics. However, the problem authors consider is different from the problem we are tackling. Authors here detect cyberbullies and cyberbullying dynamics whereas we are considering the problem of identifying cyberbullying messages in the social media data.

In most recent work, Zhong *et al.* (2016) studied detection of cyberbullying in the photo-sharing network, Instagram. The authors built an early-warning mechanism to detect images that are most vulnerable to cyberbullying attacks. The authors crawled 3,000 images from Instagram social network. The authors removed all the images that had non-english comments. The authors extracted features from comments using word2vec model, Mikolov *et al.* (2013), an offensiveness score using by finding second person pronouns in close proximity to offensive words, and bag-of-words model. The authors also use pre-trained convolutional neural networks (CNN) to extract features from images. The authors observed that, combination of all the feature sets including captions, bag-of-words, embeddings and offensiveness score gives best result.

2.2 Sentiment Classification in Social Media

Another research area related to our work is sentiment classification in social networks. Traditional sentiment classification has been extensively studied and applied

to different corpus such as product reviews Ding *et al.* (2008); Hu and Liu (2004); Liu (2012), movie reviews Pang and Lee (2004) and newspaper articles Pang *et al.* (2002). Sentiment classification can be classified into two major categories, machine learning based approach and lexicon based approaches Medhat *et al.* (2014). Machine learning based approaches involve both supervised and unsupervised sentiment classification. Various classifiers can be used to classify sentiment in a supervised way. Unsupervised sentiment classification can use emoticon information to guide overall sentiment classification model. The lexicon based approach can be further categorized into two major categories. Dictionary based approaches and corpus based approaches which uses either statistical properties or semantic properties of the classifier, respectively. Recently, sentiment analysis in social media has received increasing attention since social media is an opinion-rich resource. Sentiment analysis finds many applications in the social media realm Hu *et al.* (2013a) such as poll-rating prediction O’Connor *et al.* (2010), and event detection and prediction Bollen *et al.* (2011). However, the use of sentiment analysis to detect malicious behavior in social media is limited. One particular use of sentiment analysis to detect malicious posts in social media is done by Cambria *et al.* (2010). Xu *et al.* (2012b) uses sentiment analysis technique to identify various emotions from the bullying behavior. More specifically, the authors used a trained model and applied it to Twitter dataset to discover various emotional patterns in the bullying posts. However, this work is different from our work. We leverage the sentiment score difference between normal posts and bullying posts and proposed a sophisticated learning framework.

2.3 Abusive Language Detection in Social Media

Our work is also closely related to detecting abusive language or harassment on

the web. The first work in this area was done by Yin *et al.* (2009) which proposed a supervised classification technique using n-gram, manually developed regular expressions and contextual features. Authors also consider the abusiveness of the user’s previous messages. In this work, authors used SVM to classify the post based on the features described above. However, such methods have very limited accuracy in the social media, since the n-gram features, which accounts for the most of the feature space, is very sparse. Also, SVMs are known to work with limited accuracy in the sparse feature space settings Rendle (2010).

Another approach proposed in Sood *et al.* (2012) attempted to maintain a list of profane words in the proper context. The authors used the edit distance metric to identify the distance of words to profane words and used it to classify a sentences as abusive or normal. However, such list of profane words is difficult to maintain especially in social media where new jargon is continuously invented.

One of the first efforts to use lexical and parser features in order to detect the use of offensive language was proposed by Chen *et al.* (2012). The authors in this work do not restrict the definition of harassment. The authors used SVMs with features including n-grams, automatically derived blacklists and manual regular expression features.

Warner and Hirschberg (2012) presents one of the most comprehensive studies of hate speech detection on the web with various definitions and annotations. The authors use a similar approach to the method described above. They first derive some words that might be hateful in the dataset and finally use word sense disambiguation to determine the polarity. The methods proposed in the paper achieves an F-score of 0.63.

Finally, Djuric *et al.* (2015) uses paragraph2vec approach which is adopted from Le and Mikolov (2014) to classify clean or abusive messages. The model built by the

authors learn a low-dimensional vector representations for each social media post. The authors also demonstrate that, their model outperform simple bag-of-words based approach and several other approaches.

PROBLEM DEFINITION

In this chapter, we formally define the problem of sentiment informed cyberbullying detection. First, we present the notations we used throughout this thesis and then we formally introduce the proposed problem.

We use boldface uppercase letters (e.g., $\mathbf{A}$) to denote matrices, boldface lowercase letters to denote the vectors (e.g., $\mathbf{a}$) and lowercase letters to (e.g., a) to denote scalars. We denote the transpose of a matrix $\mathbf{A}$ as $\mathbf{A}^T$ and transpose of a vector $\mathbf{a}$ as $\mathbf{a}^T$. $\text{Tr}(\mathbf{A})$ denotes the trace of the matrix $\mathbf{A}$ if it is square. The entry of a 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 and $\|\mathbf{A}\|_F$ denotes the Frobenius norm of the matrix $\mathbf{A}$. Specifically, $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^n \sqrt{\sum_{j=1}^m \mathbf{A}_{ij}^2}$ and

Figure 3.1: Overview of the Proposed Framework

$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m \mathbf{A}_{ij}^2}$. $\|\mathbf{a}\|_2 = \sqrt{\mathbf{a}^T \mathbf{a}}$ denotes the ℓ_2 -norm of a vector.

Let $\mathbf{C} = [\mathbf{X}, \mathbf{Y}]$ denote the corpus of social media posts, where $\mathbf{X} \in \mathcal{R}^{d \times n}$ is the content matrix of these posts, $\mathbf{Y} \in \mathcal{R}^{n \times k}$ 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.*, $\mathbf{u} = \{u_1, u_2, \dots, u_m\}$, $\mathbf{R} \in \mathcal{R}^{m \times n}$ denotes the user-post relationship (as shown in Figure 3.1). Specifically, $\mathbf{R}_{ij} = 1$ if post j is posted by user u_i and $\mathbf{R}_{ij} = 0$ otherwise. Meanwhile, each post in the corpus $\mathbf{C}$ 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 score, and $\mathbf{e}$ represents the sentiment score vector for 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 relationship $\mathbf{R}$ and the sentiment score of posts $\mathbf{e}$, we aim to learn a classifier $\mathbf{W}$ to automatically detect whether unseen social media posts are normal post or bullying post.

EXPLORATORY DATA ANALYSIS

One of the important motivations of the problem we study is to investigate the correlation between sentiment information and cyberbullying behavior. To investigate the impact of the sentiment information, we perform exploratory data analysis. We first introduce two real-world social media datasets that we use in this thesis and present some observations on whether sentiment information has any impact on cyberbullying detection.

4.1 Datasets

We used two social media datasets, namely Twitter and MySpace for the problem we study. Both datasets contain labeled social media post. We employ SentiStrength Thelwall *et al.* (2010) to obtain the sentiment score for each post of Twitter and MySpace datasets.

Twitter is a microblogging website which allows users to post 140 characters messages called “Tweets”. The retweets are removed from the dataset ¹. The posts in this dataset have been manually labeled as bully or normal. This dataset has been kindly provided by Xu *et al.* (2012a).

MySpace is a social networking website which allows a registered users to view pictures, read chat and check other users’ profile information. The MySpace dataset ² used in the experiments is crawled from MySpace’s groups feature. Each post in the

¹<http://research.cs.wisc.edu/bullying/data.html>

²<http://chatcoder.com/DataDownload>

dataset is manually labeled as normal or bully. This dataset has been provided by Bayzick *et al.* (2011).

Detailed datasets statistics are summarized in Table 4.1.

Table 4.1: 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

4.2 Verifying Sentiment Score Distribution

We conduct an empirical study to verify that the sentiment distribution of the normal posts is different from the bullying posts or not. Particularly, we obtain sentiment score for each post by employing SentiStrength Thelwall *et al.* (2010) tool. Pang and Lee Pang *et al.* (2002) has verified that machine learning based methods have good prediction performance on sentiment classification task. Once we learn the model based on Thelwall *et al.* (2010), 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 4.1 shows the sentiment score distribution of the normal and the bullying posts. In Figure 4.1, the X-axis shows the sentiment polarity score and Y-axis shows the density of users. From the Figure 4.1 we can observe that two distributions are centered around different mean values. This

Figure 4.1: Sentiment Score Distribution of Normal Posts and Bullying Posts in the Twitter Dataset.

suggests that there is a clear difference between the sentiment of the normal posts and the bullying posts, and bullying posts tend to have more negative sentiment than normal posts. The sentiment distribution pattern is similar in MySpace dataset.

4.3 Verifying Sentiment Consistency

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

Let $d(p_i, p_j)$ denote the sentiment similarity distance of two social media posts p_i and p_j , which can be computed by RBF kernel based similarity: $d(p_i, p_j) = \exp(\frac{-\|e_i - e_j\|_2^2}{\sigma^2})$, where e_i and e_j are the sentiment scores of these posts, and σ denotes

the variance of the RBF kernel, usually set to be 1.

Let $\mathbf{s}_c$ and $\mathbf{s}_d$ be two vectors of the equal number of elements, each element of the first vector $\mathbf{s}_c$ denotes the sentiment similarity score of two posts p_i and p_j with the same class labels, *i.e.*, if both posts are normal or both posts are bully. Similarly, each element of the second vector $\mathbf{s}_d$ denotes the sentiment similarity score of the two posts p_i and p_j randomly chosen from the dataset. We use the two-sampled t -test to investigate whether the sentiment similarity of the two posts with the same class labels is more consistent than two randomly chosen posts. The null hypothesis is as follows: $H_0 : \tau_c \geq \tau_d$ and the alternative hypothesis is as follows: $H_1 : \tau_c < \tau_d$. Where τ_c and τ_d represent the sample means of the first group $\mathbf{s}_c$ and the second group $\mathbf{s}_d$, respectively.

The result of t -test, *i.e.*, p -values obtained on Twitter and MySpace dataset are $1.09e^{-11}$ and $1.028e^{-7}$, respectively. The results suggest 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 assumption that the sentiment scores of two posts with the same class labels are more consistent than two randomly chosen posts. The two-sampled t -test results further pave the way to incorporate sentiment information into the proposed cyberbullying detection framework.

PROPOSED FRAMEWORK - SICD

In this chapter, we introduce proposed SICD framework in detail. First, we present how to model short, noisy and unstructured post content. Then we discuss how to model user-post relationships. After that, we introduce the how to use sentiment information to enhance cyberbullying detection.

5.1 Modeling Content of Social Media Posts

In order to find better representation of text for cyberbullying detection, we employ unigram feature space with TF-IDF as feature values because, TF-IDF feature values with unigram feature space has been shown to be effective for cyberbullying detection Dadvar and De Jong (2012); Dinakar *et al.* (2011). Also, to further clean data, we perform stopwords removal and stemming.

In social media, the posts made by users are often short, noisy and unstructured. 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 cause these feature representations to be extremely sparse. Such high-dimensional and sparse feature representations result in poor prediction accuracy from traditional machine learning classifiers.

In recent years, sparse learning has been widely used to alleviate the negative effects of high-dimensional features to improve prediction performance. Hence, we employ sparse learning techniques to deal with sparse, noisy and unstructured posts.

More specifically, we use $\ell_{2,1}$ -norm of the estimator to seek a more compact feature space. The $\ell_{2,1}$ -norm regularization selects a subset of relevant features across all data instances with joint sparsity Li *et al.* (2016b). Also, we use the $\|\mathbf{W}\|_F^2$ of the estimator as another regularization term to penalize large value of coefficients in $\mathbf{W}$. The classification problem with elastic-net penalty can be learned by solving the following optimization problem:

$$\min_{\mathbf{W}} \frac{1}{2} \|\mathbf{X}^T \mathbf{W} - \mathbf{Y}\|_F^2 + \lambda \|\mathbf{W}\|_{2,1} + \frac{\lambda_A}{2} \|\mathbf{W}\|_F^2, \quad (5.1)$$

where the first term minimizes the traditional least-square loss, and λ is the parameter to control effect of the sparse regularization term. In the above objective function, the first term minimizes the least square loss between post content and class labels and the second term seeks a more compact feature representation on the content space by implicitly performing feature selection across all tasks with joint sparsity. It is worthwhile to note that we can use any other norm of the estimator ,*i.e.* ℓ_1 -norm, to perform feature selection as well. More information about various feature selection algorithms in greater detail can be found in Li *et al.* (2016a).

5.2 Modeling User-Post Relationship

Text data in social media is often linked due to the presence of various social relations. Hence, the widely adopted assumption that data are i.i.d may not be valid. Li *et al.* (2017) The texts are often correlated due to the various social relations, and these kinds of correlations can be explained by some well-received social science theories such as Homophily McPherson *et al.* (2001) and Social Influence Marsden and Friedkin (1993). In particular, we hypothesize that if two social media posts are from the same user, it is very likely that the topics of these two posts are more similar

to each other than two randomly selected posts, and thus are more likely to have same class label than two randomly chosen posts. In order to test this hypothesis, we create two equal sized vectors $\mathbf{up_c}$ and $\mathbf{up_d}$, where each element of the first vector denotes the label difference of the two posts by the same user and each element of the second vector denotes the label difference of two randomly chosen posts. We perform a two-sampled t -test to investigate the above hypothesis. We set the null hypothesis as $H_0 : \mathbf{up_c} = \mathbf{up_d}$ and the alternative hypothesis as $H_1 : \mathbf{up_c} < \mathbf{up_d}$. The t -test results, p -values, show that there is a strong evidence (with a significance level $\alpha = 0.01$) to reject the null hypothesis on both datasets.

In order to incorporate the above mentioned user-post relationships into our framework, we 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 $\mathbf{A} \in \mathcal{R}^{n \times n}$ 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:

$$\begin{aligned} & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{A}_{ij} \|\hat{\mathbf{Y}}_{i*} - \hat{\mathbf{Y}}_{j*}\|^2 \\ &= \text{Tr}(\mathbf{W}^T \mathbf{X} \mathbf{L}_\mathbf{A} \mathbf{X}^T \mathbf{W}), \end{aligned} \tag{5.2}$$

where $\hat{\mathbf{Y}} = \mathbf{X}^T \mathbf{W}$ which 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 $\mathbf{D}_\mathbf{A} = \sum_i \mathbf{A}_{ij}$.

5.3 Modeling Sentiment Information

Motivated by psychological and sociological findings on the correlation of emotions and bullying behavior, we propose to incorporate sentiment information to detect cyberbullying instances. From Chapter 4, we have a observation that sentiment

score distributions of normal posts and bullying posts are different and posts with the same labels are more likely to have more consistent sentiment score than two randomly chosen posts. Now we discuss how to leverage these observations to perform cyberbullying detection.

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

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

where $\mathcal{N}_k(e_i)$ denotes the k -nearest neighbors of post p_i in terms of sentiment correlation. Then, we propose to model the sentiment information with another Graph Laplacian Chung (1997). The key idea is if two nodes are closer to each other in the original graph $\mathbf{S}$, *i.e.*, their sentiment scores are close to each other, their labels are similar. We formulate the above idea by minimizing the following objective function:

$$\begin{aligned} & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{S}_{ij} \|\hat{\mathbf{Y}}_{i*} - \hat{\mathbf{Y}}_{j*}\|^2 \\ &= \text{Tr}(\mathbf{W}^T \mathbf{X} \mathbf{L}_{\mathbf{S}} \mathbf{X}^T \mathbf{W}), \end{aligned} \tag{5.3}$$

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 $\mathbf{D}_{\mathbf{S}} = \sum_i \mathbf{S}_{ij}$.

5.4 Sentiment Informed Cyberbullying Detection (SICD)

As illustrated from the previous section, 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 types of the information, the task of sentiment informed cyberbullying detection can be formulated as the following optimization problem:

$$\begin{aligned} \min_{\mathbf{W}} & \frac{1}{2} \|\mathbf{X}^T \mathbf{W} - \mathbf{Y}\|_F^2 + \lambda \|\mathbf{W}\|_{2,1} + \frac{\lambda_A}{2} \|\mathbf{W}\|_F^2 \\ & + \alpha \text{Tr}(\mathbf{W}^T \mathbf{X} \mathbf{L}_A \mathbf{X}^T \mathbf{W}) + \beta \text{Tr}(\mathbf{W}^T \mathbf{X} \mathbf{L}_S \mathbf{X}^T \mathbf{W}), \end{aligned} \quad (5.4)$$

where α and β are parameters to control the contribution of user-post relationship and sentiment information regularization, respectively. By solving the objective function in Eq. 5.4, results in $\mathbf{W}$, the learned classifier. To detect the cyberbullying behaviors on unseen social media post $\mathbf{x}$, we can use the following formulation:

$$\arg \max_{i \in \{bully, normal\}} \mathbf{x}^T \mathbf{W}_{*i}. \quad (5.5)$$

In next section, we introduce an efficient algorithm to solve the optimization problem in Eq. 5.4.

ALGORITHMIC DETAILS

Due to the presence of the $\ell_{2,1}$ -norm, the optimization problem in Eq. 5.4 is non-smooth but convex. In this section, we introduce an efficient algorithm to solve the optimization problem defined in Eq. 5.4 and discuss how we tackle the challenge of non-smoothness of the $\ell_{2,1}$ -norm. We also discuss the time complexity and the convergence analysis of the proposed algorithm.

6.1 Optimization Algorithm for SICD

A possible solution to solve the optimization problem in Eq. 5.4 is to use sub-gradient descent method Boyd and Vandenberghe (2004) as it is able to deal with non-smooth convex optimization problems. However, it has a very slow convergence rate, *i.e.*, $O(\frac{1}{\epsilon^2})$ where ϵ denotes the desired accuracy, and is not suitable for real-world applications.

In recent years, proximal gradient descent Ji and Ye (2009); Liu *et al.* (2009) has received increasing attention in the machine learning and data mining communities and it has been widely used in the literature to solve large-scale non-smooth convex optimization problems. It is a natural extension of the traditional gradient descent method, where the objective function to be solved can be separated into both smooth and non-smooth convex functions. In our scenario, $\|\mathbf{W}\|_{2,1}$ is the non-smooth term and the other parts form the smooth part $f(\mathbf{W})$.

$$\begin{aligned}
f(\mathbf{W}) &= \min_{\mathbf{W}} \frac{1}{2} \|\mathbf{X}^T \mathbf{W} - \mathbf{Y}\|_F^2 + \frac{\lambda_A}{2} \|\mathbf{W}\|_F^2 \\
&\quad + \alpha \text{Tr}(\mathbf{W}^T \mathbf{X} \mathbf{L}_A \mathbf{X}^T \mathbf{W}) + \beta \text{Tr}(\mathbf{W}^T \mathbf{X} \mathbf{L}_S \mathbf{X}^T \mathbf{W}). \\
F(\mathbf{W}) &= f(\mathbf{W}) + \lambda \|\mathbf{W}\|_{2,1}.
\end{aligned} \tag{6.1}$$

It is important to note that $f(\mathbf{W})$ is convex and differentiable, and the term $\|\mathbf{W}\|_{2,1}$ is non-smooth but it is still convex. In each iteration of the proximal gradient descent algorithm, the $F(\mathbf{W})$ is linearized around the current estimate of $\mathbf{W}_t$, where t indicates the t -th iteration. The value of the $\mathbf{W}$ is updated by solving following optimization problem:

$$\mathbf{W}_{t+1} = \arg \min_{\mathbf{W}} \mathcal{G}_{\eta_t}(\mathbf{W}, \mathbf{W}_t), \tag{6.2}$$

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

$$\begin{aligned}
\mathcal{G}_{\eta_t}(\mathbf{W}, \mathbf{W}_t) &= f(\mathbf{W}_t) + \langle \nabla f(\mathbf{W}_t), \mathbf{W} - \mathbf{W}_t \rangle \\
&\quad + \frac{\eta_t}{2} \|\mathbf{W} - \mathbf{W}_t\|_F^2 + \lambda \|\mathbf{W}\|_{2,1},
\end{aligned} \tag{6.3}$$

where η_t is the step size that can be determined by the backtracking line search algorithm with Armijo-Goldstein rule Bertsekas (1999). $\langle \mathbf{A}, \mathbf{B} \rangle$ denotes the dot product between two matrices $\mathbf{A}$ and $\mathbf{B}$: $\langle \mathbf{A}, \mathbf{B} \rangle = \text{Tr}(\mathbf{A}^T \mathbf{B})$. The gradient of the smooth part $f(\mathbf{W})$ in Eq. 6.1 is formulated as follows:

$$\begin{aligned}
\nabla f(\mathbf{W}_t) &= \mathbf{X} \mathbf{X}^T \mathbf{W}_t - \mathbf{X} \mathbf{Y} + \lambda_A \mathbf{W} \\
&\quad + \alpha \mathbf{X} \mathbf{L}_A \mathbf{X}^T \mathbf{W}_t + \beta \mathbf{X} \mathbf{L}_S \mathbf{X}^T \mathbf{W}_t.
\end{aligned} \tag{6.4}$$

The basic intuition behind the formulation in Eq. 6.2 is that, we can efficiently solve this problem by exploiting the structure of the $\ell_{2,1}$ -norm, hence we can achieve the same convergence rate as the gradient descent algorithm, *i.e.*, $O(\frac{1}{\epsilon})$. This convergence rate is due to the fact that we do not employ any approximation on the $\ell_{2,1}$ -norm.

In Eq. 6.3, we ignore the terms that are not related to $\mathbf{W}$ and the objective function boils down to the following optimization problem:

$$\mathbf{W}_{t+1} = \pi_{\eta_t}(\mathbf{W}_t) = \arg \min_{\mathbf{W}} \frac{1}{2} \|\mathbf{W} - \mathbf{U}_t\|_F^2 + \frac{\lambda}{\eta_t} \|\mathbf{W}\|_{2,1}, \quad (6.5)$$

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:

$$\mathbf{w}_{t+1}^i = \arg \min_{\mathbf{w}^i} \|\mathbf{w}^i - \mathbf{u}_t^i\|_2^2 + \frac{\lambda}{\eta_t} \|\mathbf{w}^i\|_2, \quad (6.6)$$

where the $\mathbf{w}_{t+1}^i$, $\mathbf{w}^i$ and $\mathbf{u}_t^i$ are the i -th row of the matrix $\mathbf{W}_{t+1}$, $\mathbf{W}_t$ 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 follows:

$$\mathbf{w}_{t+1}^i = \begin{cases} (1 - \frac{\lambda}{\eta_t \|\mathbf{u}_t^i\|_2}) \mathbf{u}_t^i; & \text{if } \|\mathbf{u}_t^i\|_2 \geq \frac{\lambda}{\eta_t}, \\ \mathbf{0}; & \text{otherwise.} \end{cases} \quad (6.7)$$

Since the proximal algorithm described above has closed form euclidean projection Liu *et al.* (2009), hence it has the same convergence rate (*i.e.*, $\frac{1}{\epsilon}$) as the traditional gradient descent algorithm for smooth convex optimization problems.

6.2 Accelerated Proximal Algorithm for SICD

In this section, we discuss how we can further accelerate the proximal algorithm to achieve an optimal convergence rate of $O(\frac{1}{\sqrt{\epsilon}})$. In particular, we employ Nesterov's method Nesterov (2013) to accelerate the proximal gradient descent in Eq. 6.2. More specifically, this accelerated algorithm is based on two sequences, $\mathbf{W}_t$ and $\mathbf{V}_t$, where $\mathbf{W}_t$ indicates the sequence of approximate solutions, and $\mathbf{V}_t$ indicates the sequence of

search points. Basically, we create a linear combination of $\mathbf{W}_t$ and $\mathbf{W}_{t+1}$ to update $\mathbf{V}_{t+1}$ as follows:

$$\mathbf{V}_{t+1} = \mathbf{W}_{t+1} + \frac{\gamma_t - 1}{\gamma_{t+1}}(\mathbf{W}_{t+1} - \mathbf{W}_t), \quad (6.8)$$

where the term $(\frac{\gamma_t - 1}{\gamma_{t+1}})$ is a linear combination coefficient. Here, $\{\gamma_t\}_{t \geq 1}$ is updated according to $\gamma_{t+1} = \frac{1 + \sqrt{1 + 4\gamma_t^2}}{2}$. In other words, the approximate solution $\mathbf{W}_{t+1}$ is computed as a “gradient” step of $\mathbf{V}_t$ through the $\mathcal{G}_{\eta_t}(\mathbf{W}, \mathbf{V}_t)$.

The whole algorithm of sentiment enhanced cyberbullying detection framework is illustrated in Algorithm 1. In particular, we use Armijo-Goldstein rule to determine the step size of the gradient descent algorithm from line 9 to 12. Finally we update γ_{t+1} according to line 15 in Algorithm 1 Liu *et al.* (2009). After running Algorithm 1 and obtaining the classifier $\mathbf{W}$, we can use $\mathbf{W}$ to detect cyberbullying behaviors on unseen test data.

6.3 Convergence Analysis

In this section, we give the convergence analysis of the proposed Algorithm 1. The convergence property of the proposed algorithm is stated in Theorem 1.

Theorem 1 *Ji and Ye (2009) Let us assume that $\{\mathbf{W}_t\}$ is the sequence of approximations obtained by the proposed Algorithm 1, then for any $t \geq 1$, we have:*

$$F(\mathbf{W}_t) - F(\mathbf{W}^*) = \frac{\leq 2\hat{L}_f \|\mathbf{W}^* - \mathbf{W}_1\|_F^2}{(t + 1)^2}, \quad (6.9)$$

where L is the Lipschitz constant w.r.t. the gradient of $f(\mathbf{W})$ in the objective function. Also, $\mathbf{W}^*$ is the solution of $F(\mathbf{W})$. More specifically, $\mathbf{W}^* = \arg \min_{\mathbf{W}} F(\mathbf{W})$.

Algorithm 1 Sentiment Informed Cyberbullying Detection in Social Media

Input: $\{\mathbf{X}, \mathbf{Y}, \mathbf{R}, \mathbf{e}, \mathbf{V}_1, \eta_1, \lambda, \alpha, \beta\}$

Output: $\mathbf{W}$

- 1: Initialize $\gamma_0 = 0, \gamma_1 = 1, t = 1$
 - 2: Initialize $\mathbf{W}_1 = \mathbf{W}_0 = \mathbf{V}_1$.
 - 3: Compute $\mathbf{A} = \mathbf{R}^T \mathbf{R}$
 - 4: Construct sentiment affinity matrix $\mathbf{S}$ from $\mathbf{e}$.
 - 5: Construct Laplacian matrix $\mathbf{L}_\mathbf{A}$ and $\mathbf{L}_\mathbf{S}$ from $\mathbf{A}$ and $\mathbf{S}$ respectively.
 - 6: Set $\nabla f(\mathbf{W}_t) = \mathbf{X}\mathbf{X}^T \mathbf{W}_t - \mathbf{X}\mathbf{Y} + \lambda_A \mathbf{W}_t + \alpha \mathbf{X}\mathbf{L}_\mathbf{A} \mathbf{X}^T \mathbf{W}_t + \beta \mathbf{X}\mathbf{L}_\mathbf{S} \mathbf{X}^T \mathbf{W}_t$.
 - 7: **while** not convergent **do**
 - 8: Set $\mathbf{U}_t = \mathbf{W}_t - \frac{1}{\eta_t} \nabla f(\mathbf{W}_t)$.
 - 9: **while** $F(\mathbf{W}_t) > \mathcal{G}_{\eta_t}(\mathbf{U}_t, \mathbf{W}_t)$ **do**
 - 10: Set $\eta_t = 2 * \eta_t$.
 - 11: Set $\mathbf{U}_t = \mathbf{W}_t - \frac{1}{\eta_t} \nabla f(\mathbf{W}_t)$.
 - 12: **end while**
 - 13: Set $\eta_{t+1} = \eta_t$
 - 14: Compute $\mathbf{W}_{t+1} = \arg \min_{\mathbf{W}} \mathcal{G}_{\eta_t}(\mathbf{W}, \mathbf{V}_t)$.
 - 15: Set $\gamma_{t+1} = \frac{1 + \sqrt{1 + 4\gamma_t^2}}{2}$
 - 16: Compute $\mathbf{V}_{t+1} = \mathbf{W}_{t+1} + \frac{\gamma_t - 1}{\gamma_{t+1}} (\mathbf{W}_{t+1} - \mathbf{W}_t)$
 - 17: Set $t = t + 1$
 - 18: **end while**
 - 19: Set $\mathbf{W} = \mathbf{W}_{t+1}$.
-

The detailed proof the theorem can be found in Ji and Ye (2009). The proof shows that the proximal algorithm, when solved using Nesterov’s accelerated method, achieves the optimal convergence rate of $O(\frac{1}{\sqrt{\epsilon}})$.

6.4 Time-Complexity Analysis

In this section, we give the time complexity analysis of the Algorithm 1. Given a corpus of $\mathbf{C}$ with n social media posts and a feature dimension of d , it requires $O(nd)$ operations in solving the least-squares formulation. The euclidean projection for the $\ell_{2,1}$ -norm according to Eq. 6.7 requires $O(2n)$ operations Liu *et al.* (2009). Third, the Laplacian regularization for the modeling of user-post relationships requires $O(nd)$. Similarly, the Laplacian regularization for the sentiment information modeling also requires $O(nd)$. Also, by employing the Nesterov’s accelerated method, we can achieve optimal convergence rate of $O(\frac{1}{\sqrt{\epsilon}})$. Hence, the total time complexity of the proposed Algorithm is $O(\frac{1}{\sqrt{\epsilon}}(nd + 2n + nd + nd)) = O(\frac{1}{\sqrt{\epsilon}}(nd))$.

EXPERIMENTS

In this chapter, we empirically evaluate the proposed SICD framework to detect cyberbullying behavior on unseen social media posts. Through performing the experiments we want to answer the following question: “How effective is the proposed SICD framework in detecting cyberbullying compared to other cyberbullying detection methods?”. We evaluate our framework on both the datasets described in the Chapter 4. After introducing the details of the experiments, we describe the effects of sentiment information and parameters in detail.

7.1 Experimental Settings

Standard experimental settings used in the literature Dadvar and De Jong (2012) have been used to evaluate cyberbullying detection methods. We apply several baseline cyberbullying detection methods on both Twitter and MySpace datasets. To avoid the bias brought by imbalanced class distributions, we use AUC and F1-measure as performance 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. 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. 5.3 using cross-validation.

Table 7.1: The F1-score and AUC score of Different Methods on Twitter Data with Varying Training Data Sizes.

	Training Ratio	20%	40%	60%	80%
F1-measure	LS	0.4105	0.4264	0.4519	0.4662
	Lasso	0.5254	0.5783	0.5927	0.6120
	MF	0.5197	0.5819	0.5916	0.6008
	USER	0.5279	0.5864	0.6023	0.6191
	POS	0.5190	0.5805	0.5939	0.6178
	SICD	0.5965	0.6265	0.6445	0.6894
AUC measure	LS	0.6142	0.6259	0.6338	0.6435
	Lasso	0.6934	0.7234	0.7318	0.7617
	MF	0.6745	0.7281	0.7335	0.7397
	USER	0.6915	0.7310	0.7426	0.7583
	POS	0.6867	0.7295	0.7431	0.7516
	SICD	0.7369	0.7684	0.7934	0.8051

7.2 Performance Evaluation

In this section, we compare our proposed SICD framework with other baseline methods and attempt to answer the question above. In order to avoid the bias brought by the size of the training data, we perform experiments with different portions of training data. We compare our proposed framework with the following methods:

- **LS**: This is traditional Least Squares method Hastie *et al.* (2001) to perform supervised classification which assumes data is i.i.d.
- **Lasso**: A supervised sparse learning framework Hastie *et al.* (2001) which uses ℓ_1 -norm of the estimator as the regularization term.

Table 7.2: The F1-score and AUC score of Different Methods on MySpace Data with Varying Training Data Sizes.

	Training Ratio	20%	40%	60%	80%
F1-measure	LS	0.4112	0.4432	0.4807	0.5018
	Lasso	0.4338	0.4759	0.4925	0.5120
	MF	0.4427	0.4917	0.5164	0.5364
	USER	0.4502	0.4818	0.5012	0.5286
	POS	0.4478	0.4789	0.4981	0.5256
	SICD	0.5086	0.5572	0.5791	0.6071
AUC measure	LS	0.6138	0.6284	0.6408	0.6516
	Lasso	0.6219	0.6378	0.6501	0.6587
	MF	0.6276	0.6495	0.6573	0.6748
	USER	0.6314	0.6418	0.6509	0.6657
	POS	0.6303	0.6392	0.6487	0.6625
	SICD	0.6549	0.6915	0.7224	0.7404

- **MF**: Since the users' posts focus on several topics Blei *et al.* (2003), the natural choice to model social media content is to use matrix factorization. For social media data, features are often the counts of words or tags, and negative values of such features are meaningless. Hence, we employ non-negativity constraint. Particularly, we use the model proposed by Lee and Seung (2001) to obtain the low-dimensional representation of the content.
- **POS**: This method has been originally proposed by Dinakar *et al.* (2011) which uses TF-IDF features, POS-tags of the bigrams, and the list of profane words as feature sets and classifies posts using SVM.

Figure 7.1: Prominent Words for Twitter Dataset

- **USER:** This method has been originally proposed by Dadvar and De Jong (2012) which uses TF-IDF features, and user related features such as gender and age as feature sets and classifies the posts using SVM.
- **SICD:** This is our proposed framework which uses user-post relationships and sentiment information.

For the *USER* baseline, if any user in both the datasets have not provided age or gender information, we impute the age information by the mean value and gender information by the most frequent value. We perform five-fold cross-validation and report the results. We first use 80% of the data as training data and hold out the 20% of the data as the test set. The Table 7.1 and Table 7.2 summarize the results on the Twitter dataset and MySpace dataset, respectively.

We draw the following observations from these two tables:

- *SICD* consistently outperforms other baseline methods on both the datasets with the varied training sizes. We also perform a pair-wise Wilcoxon signed-rank test Demšar (2006) between *SICD* and other baselines. From the test results we can

conclude that our proposed framework is significantly better (with significance level $\alpha = 0.01$) than other baselines.

- *Lasso* and *MF* both achieves better performance than *LS*. This indicates that dimensionality reduction of the original content matrix helps to improve the prediction performance. This denotes the excellent modeling of the content information.
- *SICD* achieves better results compared to other baselines with about 50% of the training data. This shows that both user-post relationships and the sentiment information help to improve the prediction performance over the pure content based methods.
- As we increase the training size from 20% to 80%, the performance of *SICD* tends to increase gradually. Which shows that more training data helps to 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 help improve prediction performance. However, the performance gain is not significant and our initial observation that users in social media intentionally obfuscate the abusive words or use intentional evasive expressions to avoid caught by list of profane words is verified in the experiments.
- Similarly, *USER* outperforms *LS*, *Lasso* and *MF* which indicates that adding user based features such as gender and age helps to improve the classification performance. However, *SICD* outperforms *USER* because the age or gender information is often scarcely available in social media due to privacy reasons Burger *et al.* (2011); Mislove *et al.* (2010); Zafarani and Liu (2013). Also, both *POS* and *USER* uses SVM as classifier to classify the post. As noted previously,

the feature space for social media data is usually sparse and SVMs usually tend to work with the limited accuracy on the sparse feature spaces Rendle (2010).

- Figure 7.1 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 the Figure 7.1. 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 Figure 7.1. As we can observe from the Figure 7.1, there is a significant overlap of the words in ground truth and prediction by *SICD* which visually demonstrates the excellent prediction quality by our proposed framework. Also, we have similar observations for the MySpace dataset.

To summarize, *SICD* outperforms other five baselines on the two real-word datasets by using sentiment information and user-post relationship information.

7.3 Impact of Sentiment Information

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

- **Content:** This is the traditional Least Squares based methods. We only use content information $\mathbf{X}$ to learn a supervised least square model.
- **Sentiment:** We first compute the sentiment score of the each post and determine its distance from the mean of the bully and normal posts group. The post is classified into the group with the shorter distance.

	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_up	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

Table 7.3: Impact of Sentiment Information in Twitter and MySpace Datasets

- **Content_up:** This method is our proposed method, where we remove the sentiment regularization part from the formulation and solve the resultant optimization problem with only user-post relationship.
- **Content_Sent:** This method is the variant of the proposed framework, where we remove the user-post regularization and solve the resultant optimization problem with only sentiment information.
- **SICD:** This method is the proposed framework that leverages both user-post relationship and sentiment information.

The experimental results for the Twitter and MySpace are summarized in the Table 7.3.

- With all the types of the information considered, *SICD* achieve the best performance compared to other baselines. The results demonstrate that our framework utilizes the different sources of information successfully to perform cyberbullying detection.
- The *Sentiment* method achieves the worst performance compared to other baselines. This observation indicates that, we can not only rely on the sentiment

information. Although, we observe the difference in the sentiment score of the normal post and bully post, we cannot just use this information to detect cyberbullying. *Content* method achieves better performance compared to *Sentiment*. It suggests that, content information is the most effective source of information to perform cyberbullying detection.

- The *Content_up* and the *Content_Sent* achieves better performance than *Content* and *Sentiment*. This result indicates that, integration of either user-post relationship or sentiment information can be used to achieve better performance compared to traditional text based cyberbullying detection methods Zafarani and Liu (2013).

In summary, the use of the sentiment information helps to improve the performance of cyberbullying detection. Sentiment Information alone is not a reliable source of information to accurately detect cyberbullying instances. With the integration of all different types of information, proposed SICD framework achieves the best performance results in detecting cyberbullying instances.

7.4 Parameter Sensitivity

Our proposed SICD framework defined in Chapter 5 has two important parameters: α and β . The parameter α and β control the contribution of user-post relationship and sentiment information, respectively. In order 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 Twitter dataset. The classification results w.r.t. AUC are shown in Figure 7.2.

It can be observed from the Figure 7.2 that when the value of β is 0.01, SICD achieves the best performance. SICD achieves relatively good performance when $\alpha < 1$

Figure 7.2: Impact of the User-Post Relationship (α) and Sentiment (β) on the Proposed Framework in Twitter Dataset

and $\beta < 1$. As the values of these two parameters increase, the performance of the proposed framework tends to decline gradually. In order to achieve good performance, the parameters α and β should be in the range $[0.001, 1]$. As observed from Figure 7.2, the results are not very sensitive to α and β . The Figure 7.2 summarizes the results on the Twitter dataset. Also, we have similar observations for the MySpace dataset.

DISCUSSIONS AND FUTURE WORK

In this work, we study a novel problem of sentiment informed cyberbullying detection in social media. The unique characteristics of the social media data and intentional obfuscation of abusive words present a unique set of challenges for cyberbullying detection in social media. Motivated by the psychological and sociological findings, we propose to leverage the sentiment information to help detect cyberbullying instances in social media. With SICD, we take a systematic approach as follows:

1. We perform an exploratory data analysis on the Twitter and MySpace datasets to investigate the effectiveness of the sentiment information. Based on these observations, we conclude that sentiment information can be a useful signal for identifying messages exhibiting cyberbullying behavior.
2. Methodologically, we incorporate sentiment information as a Laplacian regularization into our proposed sparse learning framework.
3. With variety of experiments on Twitter and MySpace datasets, we demonstrate that SICD is effective in detecting cyberbullying messages.

Unlike other approaches to cyberbullying detection, our method takes a holistic approach, using not just the actual content but also sentiment information and user-post relationship information. Our method distinguishes itself from other methods in the following ways:

1. SICD takes into account sentiment information which goes beyond just modeling language to detect cyberbullying instances.

2. We systematically incorporate sentiment information and user-post relationship information into a learning framework unlike other ad-hoc mechanisms.

It is worthwhile to note that our framework can incorporate any sentiment analysis framework, not just the one described in this thesis. Also, our method can be applied to any social media datasets to detect cyberbullying instances in an effective manner.

Our learning framework can be deployed within any social network to help social media services better detect cyberbullying instances. Also, our framework can co-exist with other traditional mechanisms to detect abuse such as regular expression detection, edit distance mechanisms etc.

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. Sometimes abusiveness in bullying posts can be across multiple posts. Another possible direction is to investigate the impact of the sarcasm information of the post in order to better detect cyberbullying behaviors.

REFERENCES

- Baldwin, T., P. Cook, M. Lui, A. MacKinlay and L. Wang, “How noisy social media text, how diffrent social media sources?”, in “IJCNLP”, pp. 356–364 (2013).
- Bayzick, J., A. Kontostathis and L. Edwards, “Detecting the presence of cyberbullying using computer software”, in “WebSci”, (2011).
- Bertsekas, D. P., “Nonlinear programming”, Athena scientific (1999).
- Blei, D. M., A. Y. Ng and M. I. Jordan, “Latent dirichlet allocation”, Journal of machine Learning research **3**, Jan, 993–1022 (2003).
- Bollen, J., H. Mao and A. Pepe, “Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.”, ICWSM **11**, 450–453 (2011).
- Boyd, S. and L. Vandenberghe, *Convex optimization* (Cambridge university press, 2004).
- Burger, J. D., J. Henderson, G. Kim and G. Zarrella, “Discriminating gender on twitter”, in “Proceedings of the Conference on Empirical Methods in Natural Language Processing”, pp. 1301–1309 (Association for Computational Linguistics, 2011).
- Cambria, E., P. Chandra, A. Sharma and A. Hussain, “Do not feel the trolls”, ISWC, Shanghai (2010).
- Card, N. A. and E. V. Hodges, “Peer victimization among schoolchildren: Correlations, causes, consequences, and considerations in assessment and intervention.”, School psychology quarterly **23**, 4, 451 (2008).
- Chen, Y., Y. Zhou, S. Zhu and H. Xu, “Detecting offensive language in social media to protect adolescent online safety”, in “Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)”, pp. 71–80 (IEEE, 2012).
- Chung, F. R., *Spectral graph theory*, vol. 92 (American Mathematical Society, 1997).
- Dadvar, M. and F. De Jong, “Cyberbullying detection: a step toward a safer internet yard”, in “Proceedings of the 21st International Conference on World Wide Web”, pp. 121–126 (ACM, 2012).
- Dadvar, M., D. Trieschnigg, R. Ordelman and F. de Jong, “Improving cyberbullying detection with user context”, in “Advances in Information Retrieval”, pp. 693–696 (2013).
- Demšar, J., “Statistical comparisons of classifiers over multiple data sets”, Journal of Machine learning research **7**, Jan, 1–30 (2006).

- Dinakar, K., B. Jones, C. Havasi, H. Lieberman and R. Picard, “Common sense reasoning for detection, prevention, and mitigation of cyberbullying”, *ACM Transactions on Interactive Intelligent Systems (TiiS)* **2**, 3, 18 (2012).
- Dinakar, K., R. Reichart and H. Lieberman, “Modeling the detection of textual cyberbullying.”, in “The Social Mobile Web”, pp. 11–17 (2011).
- Ding, X., B. Liu and P. S. Yu, “A holistic lexicon-based approach to opinion mining”, in “Proceedings of the 2008 international conference on web search and data mining”, pp. 231–240 (ACM, 2008).
- Djuric, N., J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic and N. Bhamidipati, “Hate speech detection with comment embeddings”, in “Proceedings of the 24th International Conference on World Wide Web”, pp. 29–30 (ACM, 2015).
- Hastie, T., R. Tibshirani and J. Friedman, “The elements of statistical learning”, (2001).
- Hu, M. and B. Liu, “Mining and summarizing customer reviews”, in “Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining”, pp. 168–177 (ACM, 2004).
- Hu, X., J. Tang, H. Gao and H. Liu, “Unsupervised sentiment analysis with emotional signals”, in “Proceedings of the 22nd International Conference on World Wide Web”, pp. 607–618 (ACM, 2013a).
- Hu, X., L. Tang, J. Tang and H. Liu, “Exploiting social relations for sentiment analysis in microblogging”, in “Proceedings of the sixth ACM International Conference on Web Search and Data Mining”, (2013b).
- Ji, S. and J. Ye, “An accelerated gradient method for trace norm minimization”, in “Proceedings of the 26th annual international conference on machine learning”, pp. 457–464 (ACM, 2009).
- Kokkinos, C. M. and E. Kipritsi, “The relationship between bullying, victimization, trait emotional intelligence, self-efficacy and empathy among preadolescents”, *Social psychology of education* **15**, 1, 41–58 (2012).
- Le, Q. V. and T. Mikolov, “Distributed representations of sentences and documents.”, in “ICML”, vol. 14, pp. 1188–1196 (2014).
- Lee, D. D. and H. S. Seung, “Algorithms for non-negative matrix factorization”, in “Advances in neural information processing systems”, pp. 556–562 (2001).
- Li, J., K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang and H. Liu, “Feature selection: A data perspective”, *arXiv preprint arXiv:1601.07996* (2016a).
- Li, J., X. Hu, J. Tang and H. Liu, “Unsupervised streaming feature selection in social media”, in “Proceedings of the 24th ACM International on Conference on Information and Knowledge Management”, pp. 1041–1050 (ACM, 2015).

- Li, J., X. Hu, L. Wu and H. Liu, “Robust unsupervised feature selection on networked data”, in “Proceedings of the 2016 SIAM International Conference on Data Mining”, pp. 387–395 (SIAM, 2016b).
- Li, J., L. Wu, O. Zaiane and H. Liu, “Toward personalized relational learning”, in “Proceedings of the 2017 SIAM International Conference on Data Mining”, (SIAM, 2017).
- Lieberman, H., K. Dinakar and B. Jones, “Let’s gang up on cyberbullying”, *Computer* **44**, 9, 93–96 (2011).
- Liu, B., “Sentiment analysis and opinion mining”, *Synthesis lectures on human language technologies* **5**, 1, 1–167 (2012).
- Liu, J., S. Ji and J. Ye, “Multi-task feature learning via efficient l_2, l_1 -norm minimization”, in “Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence”, pp. 339–348 (AUAI Press, 2009).
- Marsden, P. V. and N. E. Friedkin, “Network studies of social influence”, *Sociological Methods & Research* **22**, 1, 127–151 (1993).
- Mayer, J. D., R. D. Roberts and S. G. Barsade, “Human abilities: Emotional intelligence”, *Annual Review of Psychology* **59**, 507–536 (2008).
- Mckenna, J. and J.-A. Webb, “Emotional intelligence”, *British Journal of Occupational Therapy* **76**, 12, 560–561 (2013).
- McPherson, M., L. Smith-Lovin and J. M. Cook, “Birds of a feather: Homophily in social networks”, *Annual review of sociology* pp. 415–444 (2001).
- Medhat, W., A. Hassan and H. Korashy, “Sentiment analysis algorithms and applications: A survey”, *Ain Shams Engineering Journal* **5**, 4, 1093–1113 (2014).
- Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado and J. Dean, “Distributed representations of words and phrases and their compositionality”, in “Advances in neural information processing systems”, pp. 3111–3119 (2013).
- Mislove, A., B. Viswanath, K. P. Gummadi and P. Druschel, “You are who you know: inferring user profiles in online social networks”, in “Proceedings of the third ACM international conference on Web search and data mining”, pp. 251–260 (ACM, 2010).
- Nesterov, Y., *Introductory lectures on convex optimization: A basic course*, vol. 87 (2013).
- Nishina, A. and A. D. Bellmore, “When might aggression, victimization, and conflict matter most?: Contextual considerations”, *J. of Early Adolescence* pp. 5–26 (2010).
- Nishina, A. and J. Juvonen, “Daily reports of witnessing and experiencing peer harassment in middle school”, *Child development* **76**, 2, 435–450 (2005).

- Nylund, K., A. Bellmore, A. Nishina and S. Graham, “Subtypes, severity, and structural stability of peer victimization: what does latent class analysis say?”, *Child development* **78**, 6, 1706–1722 (2007).
- O’Connor, B., R. Balasubramanyan, B. R. Routledge and N. A. Smith, “From tweets to polls: Linking text sentiment to public opinion time series.”, in “ICWSM”, pp. 1–2 (2010).
- Ortony, A., G. L. Clore and M. A. Foss, “The referential structure of the affective lexicon”, *Cognitive science* **11**, 3, 341–364 (1987).
- Pang, B. and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts”, (2004).
- Pang, B., L. Lee and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques”, in “Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10”, pp. 79–86 (Association for Computational Linguistics, 2002).
- Rendle, S., “Factorization machines”, in “2010 IEEE International Conference on Data Mining”, pp. 995–1000 (IEEE, 2010).
- Singh, P., T. Lin, E. T. Mueller, G. Lim, T. Perkins and W. L. Zhu, “Open mind common sense: Knowledge acquisition from the general public”, in “OTM Confederated International Conferences” On the Move to Meaningful Internet Systems”, pp. 1223–1237 (Springer, 2002).
- Sood, S., J. Antin and E. Churchill, “Profanity use in online communities”, in “Proceedings of the SIGCHI Conference on Human Factors in Computing Systems”, pp. 1481–1490 (ACM, 2012).
- Squicciarini, A., S. Rajtmajer, Y. Liu and C. Griffin, “Identification and characterization of cyberbullying dynamics in an online social network”, in “Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015”, pp. 280–285 (ACM, 2015).
- Sui, J., *Understanding and fighting bullying with machine learning*, Ph.D. thesis, UNIVERSITY OF WISCONSIN–MADISON (2015).
- Thelwall, M., K. Buckley, G. Paltoglou, D. Cai and A. Kappas, “Sentiment strength detection in short informal text”, *Journal of the American Society for Information Science and Technology* **61**, 12, 2544–2558 (2010).
- Warner, W. and J. Hirschberg, “Detecting hate speech on the world wide web”, in “Proceedings of the Second Workshop on Language in Social Media”, pp. 19–26 (Association for Computational Linguistics, 2012).
- Xu, J.-M., K.-S. Jun, X. Zhu and A. Bellmore, “Learning from bullying traces in social media”, in “Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies”, pp. 656–666 (Association for Computational Linguistics, 2012a).

- Xu, J.-M., X. Zhu and A. Bellmore, “Fast learning for sentiment analysis on bullying”, in “Workshop on Issues of Sentiment Discovery and Opinion Mining”, p. 10 (ACM, 2012b).
- Yin, D., Z. Xue, L. Hong, B. D. Davison, A. Kontostathis and L. Edwards, “Detection of harassment on web 2.0”, Proceedings of the Content Analysis in the WEB **2**, 1–7 (2009).
- Zafarani, R. and H. Liu, “Connecting users across social media sites: a behavioral-modeling approach”, in “Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining”, pp. 41–49 (ACM, 2013).
- Zhong, H., H. Li, A. Squicciarini, S. Rajtmajer, C. Griffin, D. Miller and C. Caragea, “Content-driven detection of cyberbullying on the instagram social network”, (2016).