The Pennsylvania State University

The Graduate School

College of Information Sciences and Technology

DETECTION AND DYNAMICS OF CYBERBULLYING IN ONLINE SOCIAL NETWORKS

A Thesis in

Information Sciences and Technology

by

Yuxuan Liu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2015

The thesis of Yuxuan Liu was reviewed and approved* by the following:

Anna Cinzia Squicciarini
Assistant Professor of Information Sciences and Technology
Thesis Adviser

Erika Poole
Assistant Professor of Information Sciences and Technology

Irene Petrick
Senior Lecturer of Information Sciences and Technology
Graduate Programs Academic Coordinator of Information Sciences and Technology

Sarah Michele Rajtmajer
Postdoctoral Fellow in Department of Mathematics

Carleen Maitland
Associate Professor of Information Sciences and Technology
Interim Associate Dean for Graduate Studies of Information Sciences and Technology

*Signatures are on file in the Graduate School.

ABSTRACT

Cyberbullying, defined as bullying perpetrated through the use of information technology, is a serious problem among children, adolescents and young adults. In this thesis, we study the detection of cyberbullies and the dynamics of cyberbullying in online social networks. In order to detect cyberbullies, we use learning models based on content features, user features and social network features. Understanding cyberbullying dynamics means predicting the effect of bullying comments on users with history of non-bullying. In the course of this study, we explore the spread of cyberbullying influence through the pairwise interactions between users. Using a dataset from MySpace, we first build a model for identifying bullies and non-bullies. We build a second model formulating the relationship between the influencer and the influenced user, predicting the dynamics by which a user with history of non-bullying turns into a bully.

Based on our experimental results, we find that content features are significant in detecting cyberbullies as well as cyberbullying dynamics. We find user features to be significant in detecting cyberbullies but not significant in detecting cyberbullying dynamics. We find social network features not to be significant in detecting cyberbullies but to be significant in detecting cyberbullying dynamics.

Our models can provide moderators and network administrators an effective way to identify cyberbullies as well as develop informed insights into the dynamics of cyberbullying.

TABLE OF CONTENTS

List of Figures……vi

List of Tables……vii

Acknowledgements……ix

Chapter 1 Introduction……1

Chapter 2 Background and Related Works……4

2.1 Studies in the Field of Social Sciences ……..4

2.2 Studies in the Field of Computer and Information Sciences …….....5

2.3 Industry Tools ……7

Chapter 3 Background Notions……..10

3.1 Decision Tree……...10

3.1.1 ID3……....11

3.1.2 C4.5……...12

3.2 Evaluation Methods……14

Chapter 4 Core Solutions…….…….…….....16

4.1 Data Description and Labeling……17

4.2 Feature Analysis……...18

4.2.1 Content Features……...19

4.2.2 User Features……26

4.2.3 Social Network Features……29

4.2.4 Other Relevant Features……33

4.3 Model Selection……33

Chapter 5 Experiments and Result……35

5.1 Experiments on the Detection of Cyberbullies……35

5.2 Experiments on the Detection of Cyberbullying Dynamics……38

Chapter 6 Conclusion……43

References……46

LIST OF FIGURES

Figure 3-1: A General Structure of Decision Tree…….....10

Figure 4-1: Work Flow for Detecting Cyberbullies……16

Figure 4-2: Work Flow for Detecting Cyberbullying Dynamics……16

Figure 4-3: Statistics of Number of Profane Words based on Comments ……19

Figure 4-4: Statistics of Number of Profane Words based on Users ……20

Figure 4-5: Statistics of Number of Second Person Pronouns based on Comments…… 21

Figure 4-6: Statistics of Number of Second Person Pronouns based on Users…….21

Figure 4-7: Statistics of Comment Sentiment based on Comments……23

Figure 4-8: Statistics of Comment Sentiment based on Users……23

Figure 4-9: Statistics of Comment Length based on Comments ……. 25

Figure 4-10: Statistics of Comment Length based on Users …….25

Figure 4-11: Statistics of Users' Gender……27

Figure 4-12: Statistics of Users' Age ……28

Figure 4-13: Statistics of Number of Comments based on Users……..29

Figure 4-14: Statistics of Elapsed Time based on Thread…….....33

Figure 5-1: The Influence of Cyberbullying in the Social Network…….....38

Figure 6-1: Visualization of the Network…….....45

LIST OF TABLES

Table 2-1: Comparison among Studies……8

Table 3-1: Confusion Matrix for Binary Classification…….……14

Table 4-1: MySpace Dataset Information……..17

Table 4-2: User Demographic Information for the MySpace Dataset……...17

Table 4-3: Labeling Information ……...18

Table 4-4: Statistics of Number of Profane Words based on Comments……..19

Table 4-5: Statistics of Number of Profane Words based on Users……..20

Table 4-6: Statistics of Number of Second Person Pronouns based on Comments……...21

Table 4-7: Statistics of Number of Second Person Pronouns based on Users……...22

Table 4-8: Statistics of Comment Sentiment based on Comments...…….23

Table 4-9: Statistics of Comment Sentiment based on Users...…… 24

Table 4-10: Statistics of Comment Length based on Comments……25

Table 4-11: Statistics of Comment Length based on Users…….……..26

Table 4-12: Statistics of Users' Gender…….……..…….27

Table 4-13: Statistics of Users' Age……..28

Table 4-14: Statistics of Number of Comments based on Users...…….……...29

Table 4-15: Statistics of Social Network Features in the MySpace Dataset……..33

Table 5-1: Experimental Results for Detecting Cyberbullies ……36

Table 5-2: Accuracy of Each Feature in Detecting Cyberbullies ……37

Table 5-3: Experimental Results for Detecting Cyberbullying Dynamics……40

Table 5-4: Accuracy of Each Feature in Detecting Cyberbullies Dynamics…….42

ACKNOWLEDGEMENTS

I would like to thank my adviser, Dr. Anna Cinzia Squicciarini, for her guidance and help. Also, I would like to thank Dr. Sarah Michele Rajtmajer for her advice and help. They both have supported me throughout my thesis work with their knowledge and patience. Without their support and help, this thesis would never have been possible.

Moreover, I would like to thank Dr. Erika Poole and Dr. Irene Petrick to serve as my committee members.

Chapter 1

Introduction

Cyberbullying, defined as using information technology to willfully and repeatedly hurt, harass or threaten others [1], is a serious problem among children, adolescents and young adults. According to recent statistics [2], 52% of young people report being victims of cyberbullying. Other recent statistics [3] indicate that among more than 10,000 young people surveyed, 28% experienced cyberbullying on Twitter and 54% on Facebook. Compared with traditional bullying, cyberbullying is especially problematic because cyberbullying is not restricted by time and space [4] and can occur more frequently and intensely [5], making it more difficult to identify and control.

Although cyberbullying occurs in the cyber environment, it can cause dire consequences in the real world. Cyberbullying can lead to frustration, depression, anger and social phobia among its victims [6]. More seriously, it can lead to suicide [7]. In 2008, a 13-year-old British boy named Sam Leeson hanged himself in his bedroom because he was bullied on the social networking site Bebo [8]. Furthering the problem, retaliatory behavior can cause victims to become bullies themselves [9].

For the above reasons, it has become necessary to investigate approaches to detect and control cyberbullying behavior in social networks.

There have been a number of studies and industry tools developed to deal with this problem. Studies in the field of social sciences [10-13] analyze the different factors involved in cyberbullying as well as how to prevent cyberbullying through educating students and reporting issues immediately. Studies [14-25] in the field of computer and

information sciences focus on using machine learning techniques to detect abusive and offensive content. Existing industry tools [26-28] allow students and parents to filter or report harmful contents or contacts. Also, these tools provide a convenient way for the involvement of parents and school officials, enabling them to better protect kids online.

In this thesis, we focus on methods for detecting cyberbullies and cyberbullying dynamics in online social networks. To detect cyberbullies, we use machine learning techniques on three types of features, namely content features, user features and social network features. Content features allow us to analyze bullies based on the characteristics of their posted comments. User features allow us to characterize bullies through their demographic information. And social network features allow us to investigate bullies through their location in the network. Detecting bullies can help moderators and network administrators directly catch bullies and take necessary action such as issuing warnings and closing accounts. We also use the same feature sets together with posting times (elapsed time between posts) to characterize cyberbullying dynamics. We regard cyberbullying dynamics as the spread of cyberbullying through the pairwise interactions between users. Detecting cyberbullying dynamics means predicting whether users will be influenced by another user's cyberbullying comment. We look into the interaction between influencers and influenced users by which a user with history of non-bullying observes a peer engaging in bullying and follows suit. Predicting the dynamics of cyberbullying can help moderators and network administrators more effectively prevent the spread of cyberbullying behavior.

We use Decision Tree C4.5 with 10-fold cross-validation based on different feature combinations to detect cyberbullies and characterize the dynamics of cyberbullying. For

detecting cyberbullies, as expected, we find that content features are the most significant and social network features are the least significant. For detecting cyberbullying dynamics, we find that both content features related to the influencer (i.e. how offensive he or she is) and social network features are significant indicators. Social network features appear to indicate the importance of the role of the bullying user in the community, and his or her visibility in the network. Apparently, our results suggest that the more the user is central and key to the network, the more he or she will influence others to follow his or her reprehensible behavior.

The structure of this paper is as follows. In Chapter 2, we discuss related works. In Chapter 3, we introduce some basic concepts related to this study. In Chapter 4, we discuss the dataset as well as the analysis of three types of features. In Chapter 5, we focus on the experiments and results for detecting cyberbullies and cyberbullying dynamics. Chapter 6 is the conclusion of the study with pointers to future work.

Chapter 2

Background and Related Works

To deal with the problem of cyberbullying, there have been studies from social sciences, studies from computer and information sciences and industry tools.

2.1 Studies in the Field of Social Sciences

Studies from social sciences analyze related factors for both bullies and victims such as their socio-demographic characteristics, their personality and their habits of using technology. Moreover, they provide some solutions for preventing cyberbullying. Kift et al. [10] discuss several types of cyberbullying such as defamation, assault and invasion of privacy. Their suggestions include educating young people to increase their awareness of the dire consequences of cyberbullying and realize the necessity to report bullying incidents to trusted adults in time. Betz et al. [11] mention that it is necessary for youth to block or delete the offensive messages and report the incident to authorities instantly. Also, parents and teachers should provide guidance to help youth overcome the fear or pressures. Mishna et al. [12] discuss the possible risk factors for people to be involved in cyberbullying, including victims, bullies and bully-victims. Their considered measures include socio-demographic factors, technology use habits and parental involvement. Sabella et al. [13] investigate the accuracy of common thinking towards cyberbullying by looking into available research. Their analyzed studies come from various perspectives such as understanding, occurrence, involvement and consequence. They propose some suggestions such as developing effective school policies; educating students, parents and school staff; and developing peer helper systems.

2.2 Studies in the Field of Computer and Information Sciences

In the field of computer and information sciences, many previous studies focus on detecting cyberbullying through content-based analysis. Yin et al. [14] study the detection of online harassment based on content features, sentiment features and contextual features. Bayzick et al. [15] design a computer software called BullyTracer to detect cyberbullying based on four content features (insult words, swear words, second person pronouns and word capitalization) using the same dataset adopted in our study. Dinakar et al. [16] also focus on detecting textual cyberbullying. They first build individual topic-sensitive classifiers, and then detect the negativity and profanity of sensitive topics (intelligence, race & culture, sexuality). Reynolds et al. [17] detect cyberbullying based on the number of bad words and the density of bad words.

Some studies detect cyberbullying through applying language models. Chen et al. [18] study how to detect offensive comments as well as potential offensive users in social networks. To realize their goal, they design a Lexical Syntactic Feature-based (LSF) language model, which takes both lexical and syntactical features into consideration to first calculate sentence offensiveness. Then they incorporate sentence offensiveness values and some users' language profile features, namely sentence styles, sentence structure and cyberbullying related content to analyze users' likelihood to post offensive comments. Kontostathis et al. [19] detect cyberbullying based on two steps. The first step is to identify the frequently used cyberbullying terms and then design related queries for cyberbullying detection based on a bag-of-words language model. The second step is to apply machine learning techniques to identify more cyberbullying-related terms and querying techniques to determine the density of cyberbullying contents. Liberman et al.

[20] use common sense reasoning to detect cyberbullying. They have designed a large knowledge base for common sense knowledge, and then analyze the comment's profanity and effect on different groups of people.

Besides content features, some studies also take user features into consideration to detect cyberbullying contents. Garcia et al. [21] detect troll profiles in Twitter. They also show that their proposed model can be used for cyberbullying detection in real-world applications. Their identified features not only include content features, but also some user features such as time of publication and geoposition. Dadvar et al. [22] also detect cyberbullying considering both content features and user features. They divide content features into content-based features (number of profane words, number of first and second person pronouns, profanity windows) and cyberbullying features (number of cyberbullying words, the length of comment) and their selected user features include history of users' activities and age. They find out that the performance of detection can be improved through adding user features for analysis.

The study conducted by Huang et al. [23] takes social network features into account for detecting cyberbullying messages. They use a corpus of Twitter messages and associated ego-networks to identify content features and social network features. Their purpose is to examine whether social network features are significant in detecting cyberbullying messages. Their identified social network features include number of nodes, number of edges, number of links, degree centrality, betweenness centrality and k-core score of a node. They find out that the detection performance can be improved by considering both content features and social network features.

Dadvar et al. [24, 25] study the methods to detect bullies themselves. For their study in 2013 [24], they try to use MCES (Multi-Criteria Evaluation System) to get insights into YouTube users' behavior as well as their characteristics from the knowledge of twelve experts in the area of cyberbullying. Their study in 2014 [25] can be regarded as a continuance of their previous study. In this study, they apply supervised machine learning models in cyberbullies detection. Then they compare three types of detection methods: the expert system, the machine learning models, and a hybrid combing the two. Their extracted features include content features, user activity features and user profile features.

2.3 Industry Tools

Some industry-driven tools are available to counter cyberbullying. PureSight [26] is a software created by PureSight Technologies Ltd. This tool can better help parents protect their children from being bullied online. This tool not only automatically filters offensive contents and identify potential harmful contacts, but also allows for personalization. For example, parents can determine the list of websites or the list of contacts that they either allow or not allow their children to see or chat with. Stopit [27] is a mobile application to deal with cyberbullying. Unlike PureSight, which prevents cyberbullying from the perspective of parental control, Stopit is designed for school students. Students who use Stopit need to input their trusted adults such as parents, relatives, teachers or school officials so that they can reach out to them immediately when they see something improper. BRIM [28] is a website-based tool for students and school officials to report and track the bullying incidents. Through this tool, students can conveniently and anonymously report bullying incidents to school officials so that school officials can deal with the bullying incidents in time.

The aim of our study is to detect both cyberbullies and cyberbullying dynamics in online social networks with consideration to content features, user features and social network features. We compare our study with some similar previous studies in the area of computer and information sciences. The detailed information is in Table 2-1:

Study	Object	Feature	Dataset
Yin et al. [14]	Contents	Content features	Kongregate, Slashdot, MySpace
Bayzick et al. [15]	Contents	Content features	MySpace
Dinakar et al. [16]	Contents	Content features	Youtube
Reynolds et al. [17]	Contents	Content features	Formspring
Chen et al. [18]	Contents	Content features	Youtube
Kontostathis et al. [19]	Contents	Content features	Formspring
Garcia et al. [21]	Contents	Content features User features	Twitter
Dadvar et al. [22]	Contents	Content features User features	Youtube
Huang et al. [23]	Contents	Content features Social Network features	Twitter
Dadvar et al. [25]	Users	Content features User features	Youtube
Our study	Users and Dynamics	Content features User features Social Network features	MySpace

Table 2-1 Comparison among Studies

From Table 2-1, it can be seen that none of the previous studies focuses on studying the dynamics of cyberbullying. Therefore, our study is an important attempt to investigate the

influence of cyberbullying behavior in the social network based on users' pairwise interactions.

Chapter 3

Background Notions

3.1 Decision Tree

A decision tree is a tree-like model for classification and prediction. It is one of the most common predictive modeling approaches used in data mining and machine learning areas. The aim of the decision tree is to create a model to predict the value of the target variable based on the input variable.

The decision tree is constructed through splitting a feature space into smaller subsets repeatedly. In the structure of decision tree, each internal node refers to a test on the attribute, each branch represents the attribute value and each leaf node, also called terminal node, refers to the classification label. The basic structure of decision tree is shown in Figure 3-1:

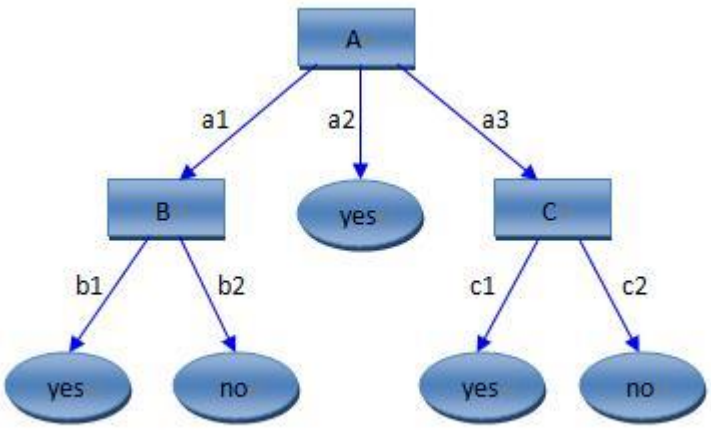


Figure 3-1: A General Structure of Decision Tree

The construction of decision tree includes two steps. The first step is building the decision tree. To build the decision tree, it starts with a root node and recursively divides the training samples into smaller subsets according to attribute value by testing for a given attribute at each node. This process continues until all the samples in one subset are the same class or the number of samples is lower than the given value. The second step is pruning the tree. The purpose of pruning the tree is to prevent overfitting so that the tree structure can become more concise and the classification accuracy can be improved [29].

The general decision tree process consists of two steps. The first step is constructing the decision tree using the training dataset with labeled categories in the recursive way. The second step is using the constructed decision tree model to classify the testing dataset.

Two notable decision tree algorithms include ID3 and C4.5.

3.1.1 ID3

ID3 (Iterative Dichotomiser 3) was developed by Ross Quinlan [30]. It is a decision tree learning algorithm based on entropy and information gain. Suppose a dataset *D* is regarded as the root node when starting to construct the tree. It first iterates through each attribute that has not been used in the set *D* and then calculates the entropy and information gain of the attribute. Next it selects the attribute that has the smallest entropy and largest information gain. After that, the set *D* is split by the selected attribute to generate the subsets. The process can recur on subsets through using the remaining attributes [31]. The process repeats until generated a decision tree that it can perfectly classify the training sample.

3.1.2 C4.5

C4.5 is a decision tree learning algorithm also developed by Ross Quinlan [32]. C4.5 is considered as an extension of ID3 based on information gain ratio. At each node of the tree, C4.5 chooses the attribute with the highest information gain ratio for decision making [33].

In order to describe the information gain ratio, we first define entropy. Entropy is a concept used to measure the degree of uncertainty in a dataset. For a set *D*, the formula to calculate entropy is as follows:

$$H(D) = -\sum_{x \in X} p(x) log_2\, p(x) \quad [31]$$

D refers to the set for which entropy is calculated; *H(D)* refers to the entropy of the set *D*; *X* refers to set of classes in set *D*; *p(x)* refers to the proportion of the number of elements in class *x* to the number of elements in set *D* [31].

For an attribute *A* in the set *D*, the formula to calculate entropy is as follows:

$$H(A, D) = -\sum_{t \in T} p(t)\, H(t) \quad [31]$$

A refers to an attribute for which entropy is calculated; *H(A,D)* refers to the entropy of the attribute *A* in set *D*; *T* refers to the subsets generated from splitting set *D* by attribute *A*; *P(t*) refers to the proportion of the number of elements in subset *t* to the number of elements in set *D*; *H(t)* refers to the entropy of subset *t* [31].

Information gain is a concept used to measure the difference in entropy from before to after the set *D* is split on the attribute *A* [31]. For an attribute *A* in set *D*, the formula to calculate information gain is as follows:

$$IG(A, D) = H(D) - H(A, D) \quad [31]$$

H(D) refers to the entropy of the set *D*; *H(A,D)* refers to the entropy of an attribute *A* in set *D* [31].

Information gain ratio refers the ratio of its information gain to its information split [34]. The bias towards multi-valued attributes can be reduced through using information gain ratio [35].

The formula to calculate information split is as follows:

$$Split\_H(A) = -\sum_{t \in T} p(t) log_2\, p(t) \quad [34]$$

A refers to an attribute for which information split is calculated*; Split_H(A)* refers to the information split of attribute *A; T* refers to the subsets generated from splitting set *D* by attribute *A*; *P(t*) refers to the proportion of the number of elements in subset *t* to the number of elements in set *D.*

The formula to calculate information gain ratio is as follows:

$$IGR(A\,, D) = \frac{IG(A\,, D)}{Split\_H(A)} \quad [34]$$

IG(A,D) refers to the information gain of the attribute *A* in set *D*; *Split_H(A)* refers to the information split of attribute *A.*

Compared with ID3, C4.5 has several improvements. Firstly, it selects attributes with information gain ratio. Secondly, it can deal with both discrete attributes and continuous attributes. Thirdly, it can handle data that contains missing values. Fourthly, it can deal with attributes that have different costs. Fifthly, it can prune the tree after creating that tree [33].

3.2 Evaluation Methods

For evaluating the efficiency of classification algorithms, frequently adopted metrics include true positive (TP) rate, false positive (FP) rate, accuracy, precision, recall and F-measure.

When using a classifier to classify instances in a particular dataset, some instances are correctly classified while some instances are wrongly classified. This can be captured in a confusion matrix. Table 3-1 shows a confusion matrix for a binary classification example:

		Predicted Class	
		C1	*C2*
True Class	*C1*	*TP*	*FN*
	C2	*FP*	*TN*

Table 3-1: Confusion Matrix for Binary Classification

In this confusion matrix, the main diagonal gives the number of positive instances that are correctly classified (*TP*) and the number of negative instances that are correctly classified (*TN*). The anti-diagonal gives the number of negative instances that are wrongly classified (*FN*) and the number of positive instances that are wrongly classified (*FP*). The actual number of positive instances $P = TP + FN$ and the actual number of negative instances $N = FP + TN$. The total number of instances in the dataset $C = P + N$.

The definition and calculation of each metric is as follows [36]:

True positive rate is defined as the proportion of correctly classified positive instances among all the positive instances. The calculation of true positive rate is given by: $TPR = TP / P$

False positive rate is defined as the proportion of wrongly classified positive instances among all the negative instances. The calculation of false positive rate is given by: $FPR = FP / N$

Accuracy (*ACC*) is defined as the proportion of correctly classified instances among all the instances. The calculation of accuracy is given by: $ACC = (TP + TN) / C$

Precision (*PREC*) is defined as the proportion of classified positive instances that are correct. The calculation of precision is given by: $PREC = TP / (TP + FP)$

Recall (*REC*) is defined as the proportion of correctly classified positive instances among all the positive instances. The calculation of recall is given by: $REC = TP / P$

F-measure (*F*) is a metric to consider both precision and recall. The calculation of F-measure is given by: $F = (PREC \times REC \times 2) / (PREC + REC)$

Chapter 4

Core Solutions

The goal of our study is to detect cyberbullies and cyberbullying dynamics in online social networks. The detection of cyberbullies is based on three feature vectors for each user's content features, demographic descriptors and social network metrics. To characterize cyberbullying dynamics, we try to predict whether a user with history of non-bullying will be influenced by another user's bullying comment and becomes a bully.

The work flow for detecting cyberbullies is shown in Figure 4-1:

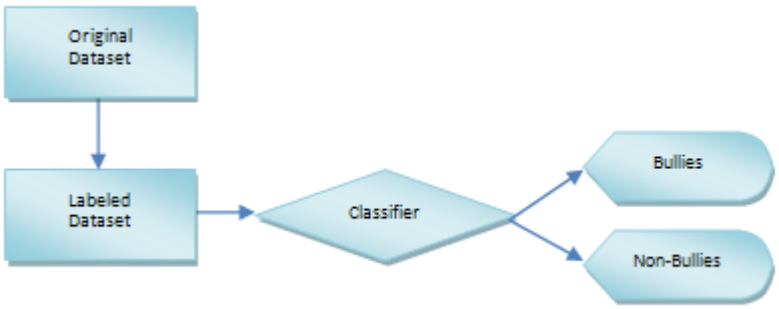


Figure 4-1: Work Flow for Detecting Cyberbullies

The work flow for detecting cyberbullying dynamics is shown in Figure 4-2:

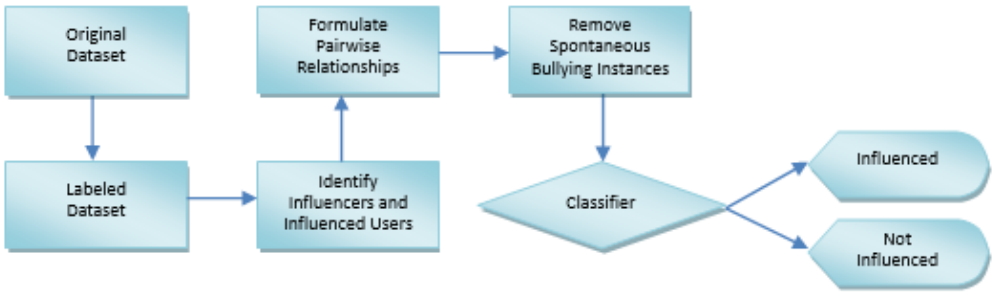


Figure 4-2: Work Flow for Detecting Cyberbullying Dynamics

4.1 Dataset Description and Labeling

The dataset we analyze is a MySpace dataset [37], containing 3,070 data records over four years in the MySpace social network. After removing repeated records, 3,032 distinct data records are obtained, with 1,129 distinct users and 118 distinct threads. The users in the MySpace dataset consist of 769 men and 358 women and their average age is 34 years old (two users do not have gender and age information). Each record includes thread id, user id, username, gender, age, location, posting time and comment body. A detailed description of the dataset is given in Table 4-1 and user demographic information is provided in Table 4-2.

Information	Total Records	Repeated Records	Remaining Records
Record	3070	38	3032
Thread	123	5	118
Distinct User	1129	0	1129
Comment	3070	38	3032

Table 4-1: MySpace Dataset Information

Information	Total	Age Blank	Age Range	Average Age
Users	1129	2	14--108	33.93
Male	769	0	15-108	33.33
Female	358	0	14-108	35.21
Gender Blank	2	2	0	0

Table 4-2: User Demographic Information for the MySpace Dataset

To determine the ground truth (whether each comment is a bullying comment or not), we have each comment hand-labeled by three labelers. We consider a particular comment to be a bullying comment if two or three of the labelers have marked it as bullying. As the result, among the 3,032 comments, 623 comments are labeled as bullying and 2,409 comments are labeled as non-bullying. Considering the perspective of an individual user, if a user posts one or more bullying comments, he or she is regarded as a bully. By this criterion, among all the 1,129 users, 384 users are labeled as bullies and 745 users are labeled as non-bullies. These results are summarized in Table 4-3:

Type	Comments	Users
Bullies	623	384
Non-bullies	2409	745
Total	3032	1129

Table 4-3: Labeling Information

Note that 623 comments are considered bullying, which accounts for about 20.55% of all the 3,032 comments, while 384 users are considered bullies, accounting for about 34.01% of all the 1,129 users. The hand-labeled data will be used as the ground truth.

4.2 Feature Analysis

In order to detect cyberbullies and characterize the dynamics of cyberbullying in online social networks, we use three sets of features for classification, namely content features, user features and social network features.

4.2.1 Content Features

For each comment, we extract four content features: the number of profane words, the number of second person pronouns, comment sentiment and length of comment.

C1: Number of profane words

This feature is used to measure the presence and frequency of offensive or inappropriate words within comments. To determine whether a word is a profane word or not, we adopt a standard profane word list [38] as the basis of judgment. We count the number of profane words for each comment. Statistics with respect to profane words are given in Figure 4-3, Table 4-4, Figure 4-4 and Table 4-5:

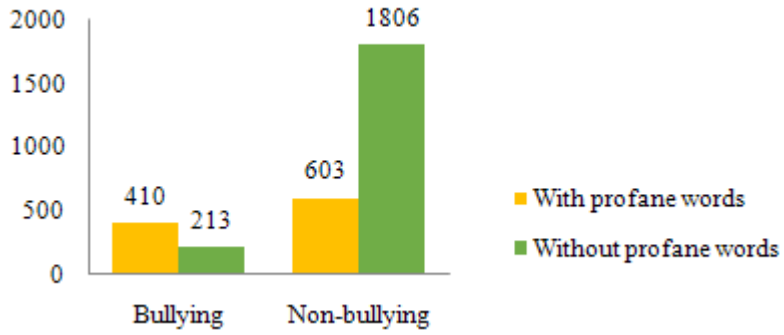


Figure 4-3: Statistics of Number of Profane Words based on Comments

Comments	Bullying	Non-bullying	Total
With profane words	410	603	1013
Without profane words	213	1806	2019
Total	623	2409	3032

Table 4-4: Statistics of Number of Profane Words based on Comments

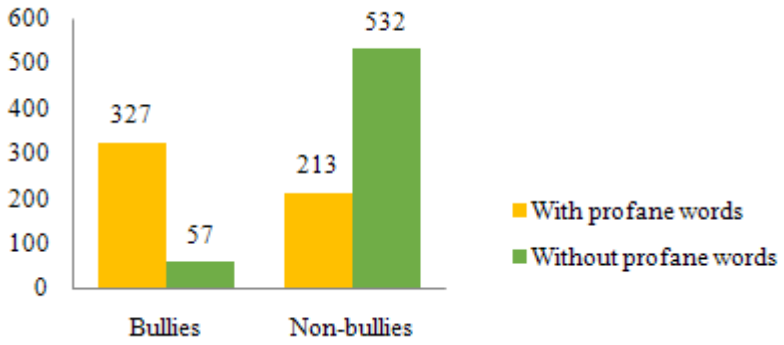


Figure 4-4: Statistics of Number of Profane Words based on Users

Users	Bullies	Non-bullies	Total
With profane words	327	213	540
Without profane words	57	532	589
Total	384	745	1129

Table 4-5: Statistics of Number of Profane Words based on Users

From Figure 4-3, Table 4-4, Figure 4-4 and Table 4-5, it can be seen that among bullies' comments, the proportion of comments with profane words is significantly larger.

C2: Number of second person pronouns

Previous research [15, 18] indicates that within the comments containing profane words, the appearance of second person pronouns will make those comments more offensive. Second person pronouns can be used to detect whether a comment is aimed at a specific person or not. To take the number of second person pronouns into consideration, we consider those comments with profane words. Statistics with respect to second person pronouns are given in Figure 4-5, Table 4-6, Figure 4-6 and Table 4-7:

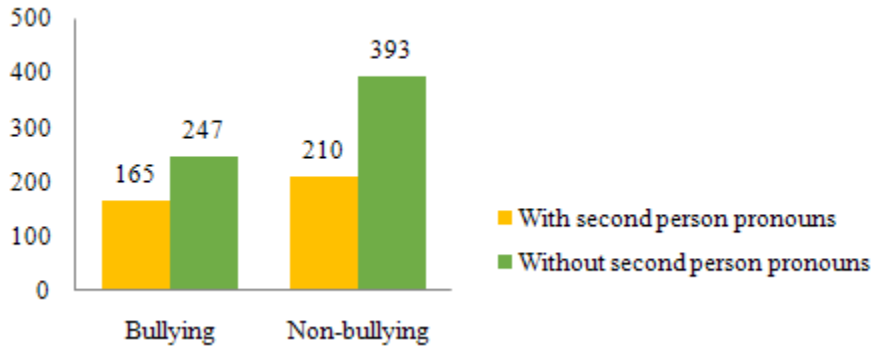


Figure 4-5: Statistics of Number of Second Person Pronouns based on Comments

Comments with Profane Words	Bullying	Non-bullying	Total
With second person pronouns	163	210	373
Without second person pronouns	247	393	640
Total	410	603	1013

Table 4-6: Statistics of Number of Second Person Pronouns based on Comments

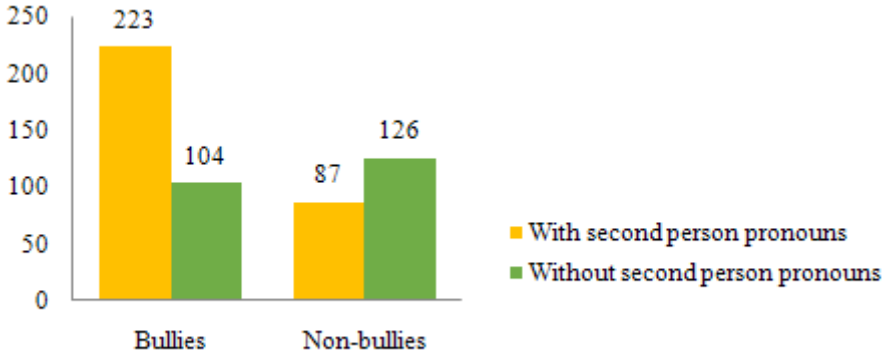


Figure 4-6: Statistics of Number of Second Person Pronouns based on Users

User with Profane Words	Bullies	Non-bullies	Total
With second person pronouns	223	87	310
Without second person pronouns	104	126	230
Total	327	213	540

Table 4-7: Statistics of Number of Second Person Pronouns based on Users

From Figure 4-5 and Table 4-6, it can be seen that from comment perspectives of this feature, there is no obvious difference between bullying comments and non-bullying comments. However, from Figure 4-6 and Table 4-7, it can be seen that from the user perspective of this feature, the number of second person pronouns can distinguish bullies and non-bullies.

C3: Comment sentiment

Sentiment analysis is commonly used in cyberbullying detection because it can help us detect offensive contents that do not contain profane words. Usually comment sentiment is used for determining the attitude of a user with respect to some topic or the overall contextual polarity of a document. In our analysis, we use a sentiment analysis tool called Semantria [39] to classify all the comments into one of the three categories: positive, negative and neutral. Statistics with respect to comment sentiment are given in Figure 4-7, Table 4-8, Figure 4-8 and Table 4-9:

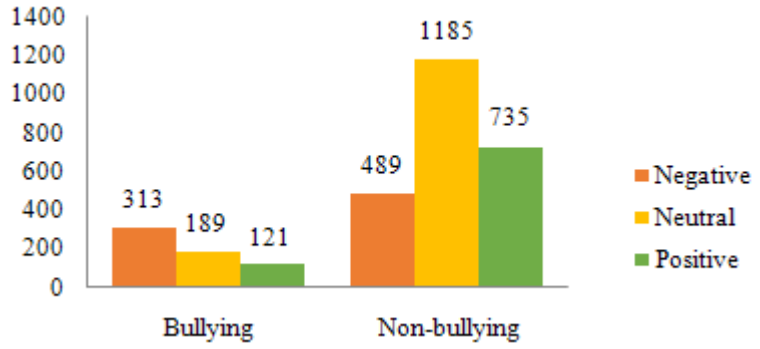


Figure 4-7: Statistics of Comment Sentiment based on Comments

Comments	Bullying	Non-bullying	Total
Negative sentiment	313	489	802
Neutral sentiment	189	1185	1374
Positive sentiment	121	735	856
Total	623	2409	3032

Table 4-8: Statistics of Comment Sentiment based on Comments

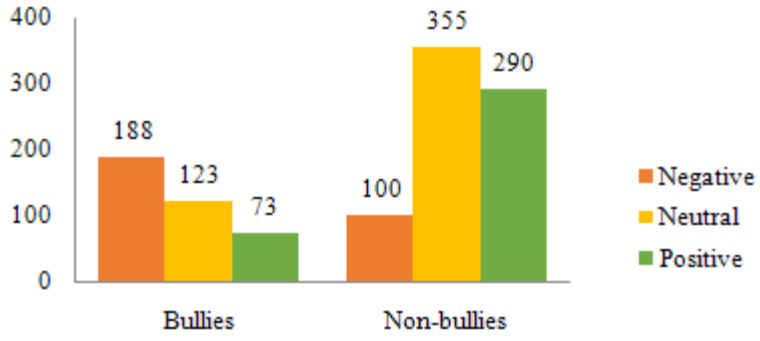


Figure 4-8: Statistics of Comment Sentiment based on Users

Users	Bullies	Non-bullies	Total
Negative sentiment	188	100	288
Neutral sentiment	123	355	478
Positive sentiment	73	290	363
Total	384	745	1129

Table 4-9: Statistics of Comment Sentiment based on Users

From Figure 4-7, Table 4-8, Figure 4-8 and Table 4-9, it can be seen that most bullies' comments tend to have negative sentiments whereas non-bullies' comments are more likely to have normal or positive sentiments.

C4: Length of comment

Previous research [18] shows that bullying comments tend to be shorter than normal comments. Therefore, length of comment can be considered as a feature for detecting bullies. The length of a comment is determined by the total number of words, regardless of space, punctuations and emoticons. The statistical results with respect to length of comment are given in Figure 4-9, Table 4-10, Figure 4-10 and Table 4-11:

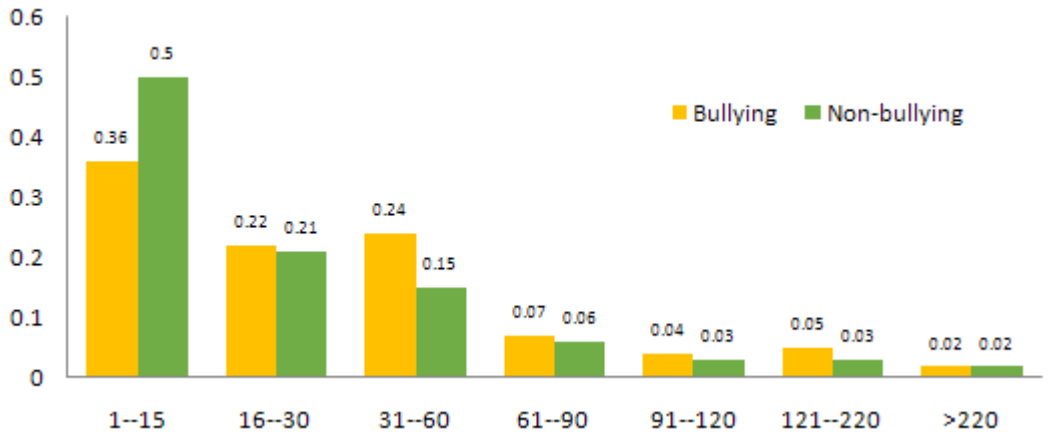


Figure 4-9: Statistics of Comment Length based on Comments

Length of Comments		1-15	16-30	31-60	61-90	91-120	121-220	>220
Bullying	Comments	220	137	152	43	26	31	14
	Ratio	0.36	0.22	0.24	0.07	0.04	0.05	0.02
Non-bullying	Comments	1209	517	350	140	61	84	48
	Ratio	0.50	0.21	0.15	0.06	0.03	0.03	0.02

Table 4-10: Statistics of Comment Length based on Comments

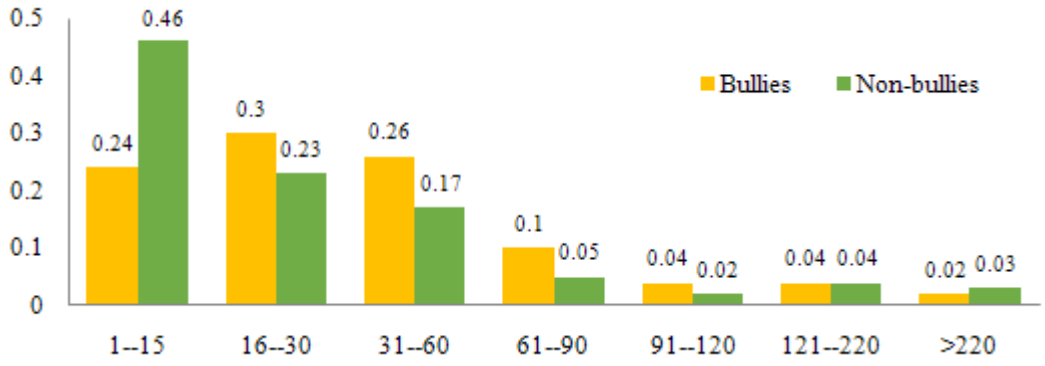


Figure 4-10: Statistics of Comment Length based on Users

Length of Comments		1-15	16-30	31-60	61-90	91-120	121-220	>220
Bullies	Users	94	115	99	37	14	16	9
	Ratio	0.24	0.30	0.26	0.10	0.04	0.04	0.02
Non-bullies	Users	336	169	126	38	17	26	19
	Ratio	0.46	0.23	0.17	0.05	0.02	0.04	0.03

Table 4-11: Statistics of Comment Length based on Users

From Figure 4-9, Table 4-10, Figure 4-10 and Table 4-11, it can be observed from both user and comment perspectives that most comments contain less than 60 words in total. When the total number of words is less than 15, the ratio of bullying comments/bullies is lower than that of non-bullying comments/non-bullies. When the total number of words is between 16 and 60, the ratio of bullying comments/bullies is higher than that of non-bullying comments/non-bullies. When the total number of words is higher than 60, there is no obvious difference between bullying comments/bullies and non-bullying comments/non-bullies.

4.2.2 User Features

For user features, we consider gender, age and the total number of comments. User features can help us identify bullies based on their demographic information and online activity.

U1: Gender

In our current dataset, we find that among bullies, the ratio of male bullies is significantly higher than that of female bullies. Statistics with respect to gender are given in Figure 4-11 and Table 4-12 (2 users do not have gender information):

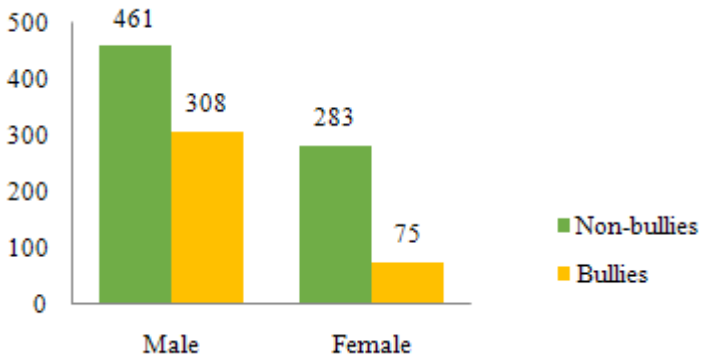


Figure 4-11: Statistics of Users' Gender

Users	Bullies	Non-bullies	Total
Male	308	461	769
Female	75	283	358
Total	383	744	1127

Table 4-12: Statistics of Users' Gender

U2: Age

Previous research [24, 25] shows that user's age is an important user feature because the frequency of cyberbullying occurrence and word use can change in different ages. These are self-declared values, and may not be truthful. However, we ignore this issue as the

reported ages are overall consistent with known averages of population of social network users. Statistics about users' age distribution are given in Figure 4-12 and Table 4-13 (2 users do not have age information):

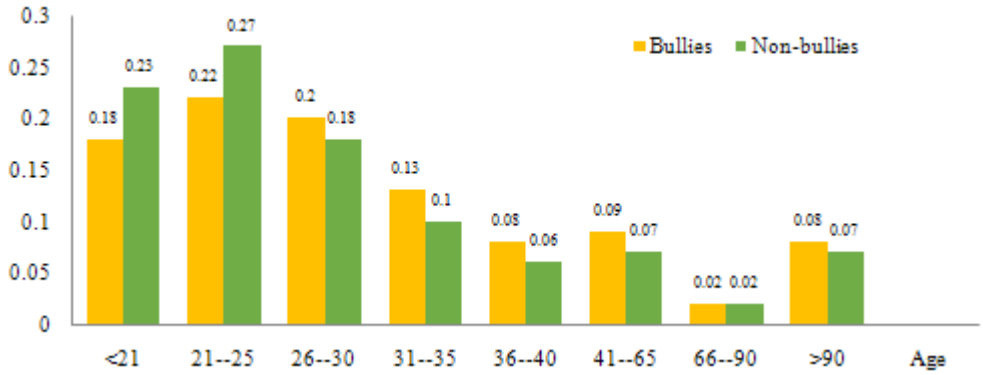


Figure 4-12: Statistics of Users' Age

Age		<21	21-25	26-30	31-35	36-40	41-65	66-90	>90
Bullies	Users	67	84	76	50	31	34	9	32
	Ratio	0.18	0.22	0.20	0.13	0.08	0.09	0.02	0.08
Non-bullies	Users	171	198	133	78	46	53	13	52
	Ratio	0.23	0.27	0.18	0.10	0.06	0.07	0.02	0.07

Table 4-13: Statistics of Users' Age

From Figure 4-12 and Table 4-13, it can be observed when the age is less than 25, the ratio of non-bullies is higher. When the age is higher than 25, the ratio of bullies is higher.

U3: Number of comments

The number of comments per user is taken to reflect a user's activity level in the social network. We notice, in our dataset, that bullies post more comments than non-bullies.

The average numbers of comments are 4.79 for bullies and 1.49 for non-bullies. These findings are detailed in Figure 4-13 and Table 4-14:

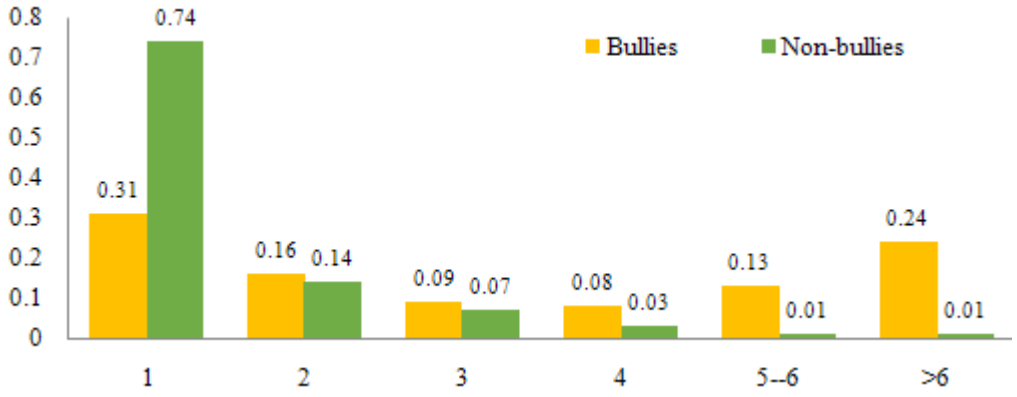


Figure 4-13: Statistics of Number of Comments based on Users

Number of Comments		1	2	3	4	5-6	>6
Bullies	Users	118	62	36	30	51	87
	Ratio	0.31	0.16	0.09	0.08	0.13	0.23
Non-bullies	Users	538	102	49	25	8	9
	Ratio	0.74	0.14	0.07	0.03	0.01	0.01

Table 4-14: Statistics of Number of Comments based on Users

4.2.3 Social Network Features

A social network is a set of nodes (or vertices) and edges. A node represents an individual and an edge represents the relationships between them.

Our set of social network features consists of five common social network graph metrics: degree centrality, closeness centrality, betweenness centrality, eigenvector centrality and

clustering coefficient. Degree centrality, closeness centrality and betweenness centrality measure the node's position and popularity in the social network. Eigenvector centrality measures the node neighbors' position and popularity in the social network. Clustering coefficient measures the mutual connections between the node's neighbors in the social network.

S1: Degree centrality

Degree centrality is defined as the number of edges that directly links to a node. If a node has more direct links with other nodes, this node is more influential in the network. The formula to calculate degree centrality is:

$$C_D^w(i) = \sum_j^N w_{ij} \qquad [40]$$

In this formula, *i* refers to the focal node; *j* refers to other nodes in the network; *N* refers to the total number of nodes and *w* represents the weighted adjacency matrix. If there is connection between node *i* and node *j*, the value of w_{ij} is the weight of the connection. If there is no connection between node *i* and node *j*, the value of w_{ij} is 0 [40].

S2: Closeness centrality

Closeness centrality is defined as the inverse of the shortest distance between a node and all other nodes reachable from this node. The higher a user's closeness centrality score is, the shorter the distance between the user and other users is so that the faster the user can affect other users. The formula to calculate closeness centrality is:

$$C_C^w(i) = \frac{N-1}{\sum_{j=1, j\neq i}^N d_{ij}} \qquad [41]$$

In this formula, *i* refers to the focal node; *j* refers to other nodes in the network; *N* refers to the total number of nodes and d_{ij} refers to the shortest distance between node *i* and node *j* [41].

S3: Betweenness centrality

Betweenness centrality of a node is defined as the portion of shortest paths between node pairs in the network that pass through that node. The higher a user's betweenness centrality score is, the more influential the user will be during the information propagation. The formula to calculate betweenness centrality is:

$$C_B^w(i) = \frac{g_{jk}(i)}{g_{jk}} \quad [40]$$

In this formula, g_{jk} refers to the number of shortest distances between two nodes and $g_{jk}(i)$ refers to the number of shortest distances that pass through node *i* [40].

S4: Eigenvector centrality

Eigenvector centrality is a measure to assign a score to each node based on the rule that connections to high-scoring nodes contribute more to the score of that node than equal connections to low-scoring nodes. The formula to calculate eigenvector centrality is:

$$C_e^w(i) = \lambda^{-1} \sum_j A_{ij}\, x_j \quad [42]$$

In this formula, λ refers to a constant; *x* refers to the eigenvector centrality score; *A* refers to the adjacency matrix of two nodes [42].

S5: Clustering coefficient

Clustering coefficient refers to the portion that a node's neighbors are connected to each other. The value of clustering coefficient is 1 if all of a node's neighbors are connected to each other. The value of clustering coefficient is 0 if none of a node's neighbors is connected to each other. The formula to calculate clustering coefficient is:

$$CC(i) = \frac{1}{2(K_i - 1)C_D^W(i)} \sum_{j,h} (w_{ij} + w_{ih})\, a_{ij} a_{ih} a_{jh} \qquad [43]$$

In this formula, k_i refers to all the neighbor nodes of node i; $C_D^W(i)$ refers to the degree centrality of node i; w refers to the weights between two nodes and a refers to the adjacency matrix [43].

To calculate these social network metrics, we construct a social network graph from our dataset. In the graph, users are considered nodes and the connections between users are considered edges. If user *A* and user *B* appear in the same thread, then we determine that there exists connection between user *A* and user *B*. The number of nodes in the graph is 1,129 and the number of edges is 15,579. Social network features can help us identify bullies through looking into their locations in the network.

Table 4-15 describes various social network metrics derived from the social network graph induced from the MySpace dataset, including degree centrality, betweenness centrality, closeness centrality, eigenvector centrality and clustering coefficient (CC) for bullies and non-bullies.

Users	Degree	Betweenness	Closeness	Eigenvector	Clustering Coefficient
Bullies	29.04	2128.74	0.34	1.04	0.40
Non-bullies	26.96	720.35	0.35	0.65	0.48

Table 4-15: Statistics of Social Network Features in the MySpace Dataset

4.2.4 Other Relevant Feature

R1: Elapsed time

Elapsed time refers to the difference between the posting times of two users. For the Figure 4-14, it can be observed that the shorter the interaction time is, the more likely that the users will be influenced by the bullies.

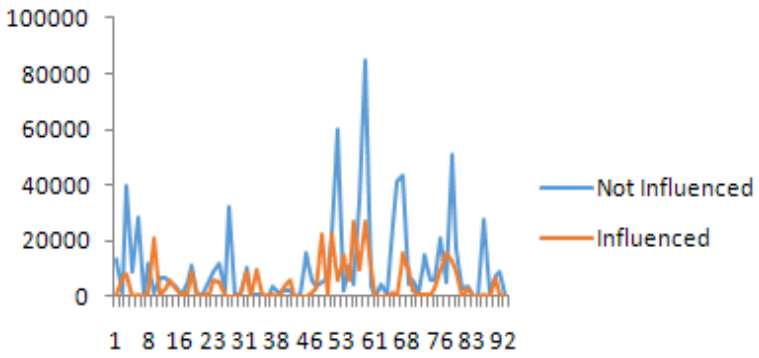


Figure 4-14: Statistics of Elapsed Time based on Thread

4.3 Model Selection

There are many machine learning algorithms for classification. Notable algorithms include decision tree, Naïve Bayes and support vector machine (SVM).

Decision tree is a predictive model that uses a tree-like model to illustrate the decisions and possible outcomes. Naïve Bayes classifier is a supervised learning algorithm based on applying Bayes' theorem with naïve assumption of independence between the features [44]. SVM is a machine learning model that tries to find a hyperplane or a set of hyperplanes that can maximize the margins between the classes [45].

There are several reasons why we apply decision tree as the basis for building the model. The first reason is that our data contains both numerical data (e.g. number of posts, length of comments, and number of profane words) and categorical data (e.g. gender). Also, some features of the data are correlated (e.g. betweenness centrality and closeness centrality). Moreover, the effectiveness of decision tree will not be negatively affected by redundant features. Naïve Bayes method is considered to have a solid mathematical foundation as well as stable classification efficiency. The assumption of Naïve Bayes classifier is that each feature is independent. However, since some of our features are correlated, Naïve Bayes is not a suitable method for our experiments. SVM is a machine learning method that can be effective in high dimensional spaces. Also, it can deal with both linear and non-linear problems. The disadvantage is that the processing time of SVM is slow. Also, to deal with non-linear problems, the kernel function is hard to determine. However, kernel function plays a pivotal role in determining the performance of SVM.

Chapter 5

Experiments and Result

For our experiments, we use the WEKA software [46], which is open source data mining software. WEKA is written in Java and includes a comprehensive collection of machine learning algorithms.

5.1 Experiments on the Detection of Cyberbullies

For detecting cyberbullies, we take three sets of features into consideration, namely content features, user features and social network features. Since content features are calculated using comments, we convert the values of content features into user perspectives. Specifically, we add up the content feature values of all the comments posted by a user to derive a total value. We then calculate the average content feature values of the user by dividing the total value by the number of comments.

We use the Decision Tree C4.5 classifier (Implemented as J48 in WEKA) with 10-fold cross-validation. We consider all the three sets of features in different combinations. This allows us to find the significance of each type of feature for classification. More specifically, we consider five different feature combinations: only content features, only user features, only social network features, both content features and user features, all the features. To evaluate the obtained result, we use true positive rate (TPR), false positive rate (FPR), F-measure and overall accuracy. The experimental results with different combinations of features are shown in Table 5-1:

Feature Combinations	Content Features	User Features	Social Network Features	Content Features & User Features	All Features
TPR	0.821	0.756	0.717	0.852	0.839
FPR	0.258	0.408	0.471	0.214	0.232
F-measure	0.817	0.734	0.685	0.849	0.836
Overall Accuracy	0.821	0.756	0.717	0.852	0.839

Table 5-1: Experimental Results for Detecting Cyberbullies

From Table 5-1, it can be observed that content features are the most significant feature. With only content features, the overall accuracy is 82.1%. User features are also significant in detecting cyberbullies. When user features and content features are combined, the overall accuracy is 85.2%, which is higher than the overall accuracy with content features only. The accuracy obtained by considering social network features only is 71.7%, which is lower than both the accuracy (82.1%) of considering content features only and the accuracy (75.6%) of considering user features only. This means social network features are not as significant as content features and user features in detecting cyberbullies. Also, when social network features are combined with content features and user features, the overall accuracy is 83.9%, which is slightly lower than the combination with content features and user features.

When we run the experiments above, we find that some features have “negative effects” and lower the overall accuracy for classification. To validate the contribution of each feature, we use each feature alone for classification and calculate the accuracy.

A more detailed analysis of each feature in detecting cyberbullies is shown in Table 5-2:

Accuracy (%)	Feature Name	Feature Type
78.12	Number of Profane Words	Content
78.03	Number of Second Person Pronouns	Content
76.26	Comment Sentiment	Content
75.47	Number of Comments	User
72.28	Clustering Coefficient	Social Network
71.83	Betweenness Centrality	Social Network
68.02	Degree Centrality	Social Network
67.49	Eigenvector Centrality	Social Network
67.23	Closeness Centrality	Social Network
65.99	Age	User
65.99	Length of Comments	Content
65.99	Gender	User

Table 5-2: Accuracy of Each Feature in Detecting Cyberbullies

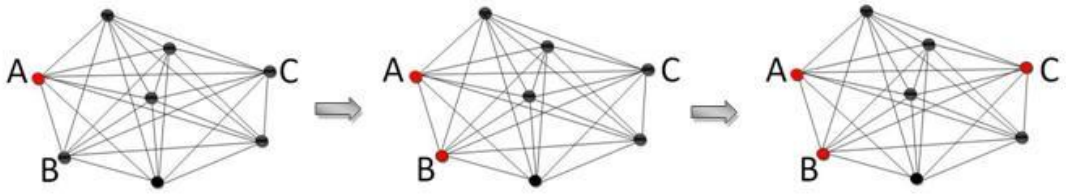
The value of the first column refers to the accuracy when the feature alone is used for classification. From Table 5-2, it can be observed that number of profane words, number of second person pronouns and comment sentiment are the most important features for detecting cyberbullies. This is not surprising, as content features are direct manifestations of bullying. Nevertheless, length of comments, along with age and gender, are the least significant compared with other features. Considering gender, the previous statistical analysis towards gender indicates that the ratio of male bullies is much higher than that of

female bullies. However, the machine learning classification result shows that gender is not significant for classify bullies and non-bullies. Social network features are also not highly ranked, which means whether user is a bully is not much relevant to his or her position and connection within the network.

5.2 Experiments on the Detection of Cyberbullying Dynamics

Cyberbullying dynamics is regarded as the spread of cyberbullying influence in the pairwise interactions between users. Detecting cyberbullying dynamics means predicting whether users will be influenced by a cyberbullying comment and will also exhibit bullying behavior. In the pairwise interactions of users, we build the relationship between the influencer and the influenced user by which a user with history of non-bullying observes bullying behavior and turns into a bully.

This is a more specific example to show how normal users are influenced by bullies. There are three users: user *A*, user *B* and user *C*. In Figure 5-1 (a), at the first time point, user *A* is a bully while user *B* and user *C* are non-bullies. In Figure 5-1 (b), at the second time point, user *B* becomes a bully. In Figure 5-1 (c), at the third point, user *C* becomes a bully.



(a) (b) (c)

Figure 5-1: The Influence of Cyberbullying in the Social Network

For the implemented dataset, we extract the pairwise users based on the thread and the posting time. In one thread, we create two-user interactions based on the sequence of posting time. For instance, in the thread *TH*, there are three users: user *A*, user *B* and user *C*. User *A* posts at 1 pm, user *B* posts at 2 pm and user C posts at 3 pm. The two-user interactions in this thread will include $A \rightarrow B$, $A \rightarrow C$ and $B \rightarrow C$.

We add the pairwise user's feature information together in one record. The influence result will be considered as positive if all the conditions below are met:

- User *A* posts a bullying comment in the thread.
- User *B* has observed history of non-bullying.
- User *B* posts a bullying comment in the thread, following the comment of User *A*.

In the implemented dataset, there are 43,969 pairwise interactions in total. Among these instances, 2,182 records are instances of influence whereas 41,787 records are instances of non-influence. Since the classes are imbalanced, we subsample the majority class, which is the class of non-influence instances. More specifically, we randomly sample 2,182 instances of non-influence and combine them with the 2,182 instances of influence to get the dataset with 4,364 instances in total. This process is repeated for 10 randomly subsampled datasets and averaged to obtain the final results.

For detecting cyberbullying dynamics, we also take content features, user features, social network features into consideration. For user features and social network features, we consider both influencer users and influenced users. For content features, we consider the content features specific to influencers' comments. Besides these three types of features, we also consider elapsed time, which refers to the time difference between two posts in the pairwise interaction.

We use the Decision Tree C4.5 classifier (Implemented as J48 in WEKA) with 10-fold cross-validation. The experiment is done with five different combinations of features. In the first combination (F1), we consider all the features. In the second combination (F2), we consider all the features except the content features specific to influencers' comments. In the third combination (F3), we consider all the features except the user features. In the fourth combination (F4), we consider all the features except the social network features. In the fifth combination (F5), we consider all the features except the elapsed time. The four metrics for evaluating the results are also true positive rate (TPR), false positive rate (FPR), F-measure and overall accuracy. The experimental results are shown in Table 5-3:

Feature Combinations	F1	F2	F3	F4	F5
TPR	0.815	0.742	0.817	0.787	0.818
FPR	0.185	0.258	0.183	0.213	0.182
F-measure	0.814	0.741	0.816	0.787	0.817
Overall Accuracy	0.815	0.742	0.817	0.787	0.818

Table 5-3: Experimental Results for Detecting Cyberbullying Dynamics

From Table 5-3, it can be observed that social network features are a significant type of features for detecting cyberbullying dynamics. Without social network features, the overall accuracy decreases to 78.7% from 81.5%. Still, content features specific to influencers' comments are the most significant in detecting cyberbullying dynamics. When content features are added, the overall accuracy improves to 81.5% from 74.2%. However, user features and elapsed time are not significant indicators. When user

features or elapsed time are removed, the overall accuracy does not have significant changes (very slight increase). The significance of content features specific to influencers is not surprising, because content features are always strong indicators of cyberbullying occurrences. The significance of social network features means the more the user is central and key to the network, the more he or she will influence others or be influenced by others. Since user features are not significant, it means the user's age, gender and number comments do not have much relation to the likelihood that he or she will influence others. Elapsed time is not significant, which means the time difference between posting comments is also not much related to user influence.

Same as the previous experiments, to validate the contribution of each feature, we use each feature alone for classification and calculate the accuracy.

A more detailed analysis of each feature in detecting cyberbullying dynamics is shown in Table 5-4 (1 means the feature is the influencer's feature; 2 means the feature is the influenced user's feature).

The value of the first column refers to the accuracy when the feature alone is used for classification. Not surprisingly, the most significant feature is the number of profane words posted by influencers. Also, all the social network features except clustering coefficient are highly ranked. This means the dynamics of cyberbullying is more related to users' position but less related to the mutual connection between the user's neighbors. Moreover, it can be observed that influencers' social network features are ranked higher that influenced users' social network features, which means when the user is more central and popular in the social network, he or she is more likely to influence others rather than be influenced by others.

Accuracy (%)	Feature Name	Feature Type
70.19	Number of Profane Words1	Content
68.50	Eigenvector Centrality1	Social Network
68.33	Comment Sentiment1	Content
67.54	Closeness Centrality1	Social Network
67.08	Betweenness Centrality1	Social Network
66.74	Eigenvector Centrality2	Social Network
64.88	Degree Centrality1	Social Network
62.99	Betweenness Centrality2	Social Network
62.60	Closeness Centrality2	Social Network
60.37	Length of Comments1	Content
60.23	Degree Centrality2	Social Network
58.64	Number of Second Person Pronouns1	Content
56.40	Clustering Coefficient2	Social Network
56.22	Number of Comments1	User
56.06	Number of Comments2	User
56.02	Elapsed Time	Other
55.84	Gender2	User
55.69	Clustering Coefficient1	Social Network
54.11	Age2	User
53.96	Gender1	User
51.50	Age1	User

Table 5-4: Accuracy of Each Feature in Detecting Cyberbullying Dynamics

Chapter 6

Conclusion

In this thesis, we present a study to detect cyberbullies and cyberbullying dynamics. More specifically, we first detect cyberbullies based on Decision Tree C4.5 classifier using content features, user features and social network features. Then we also use Decision Tree C4.5 classifier to detect the dynamics of cyberbullying with consideration to the same feature sets and the elapsed time between comments. Cyberbullying dynamics refers to the spread of cyberbullying influence in the pairwise interactions between users. Detecting cyberbullying dynamics means predicting whether users with history of non-bullying will be influenced by a bullying comment in the user interactions. In the pairwise interactions of users, we build a model formulating the relationship between the influencer and the influenced user, predicting the dynamics by which a user with history of non-bullying turns into a bully.

In the experiment for detecting cyberbullies, we find that content features are the most significant indicators whereas social network features less significant than other features, which means whether user is a bully or not is not much related to his or her position or role in the network. With respect to cyberbullying dynamics, we find that content features specific to influencers' comments and social network features are significant indicators. The significance of social network features means the user's position and role in the network is related to the user influence. However, user features and elapsed time are not significant indicators.

The novel part of our study is detecting cyberbullies as well as the dynamics of cyberbullying with consideration to various types of features. To the best of our knowledge, none of the previous studies looks into how the cyberbullying behavior spreads and influences others in the social network, our study is an important attempt to deal with this issue. We tried different feature combinations to explore the significance of each feature type in the experiments to address this question. For content features, they are significant in both detecting cyberbullies and cyberbullying dynamics. For user features, age and gender are not significant in both experiments while the number of comments plays a significant role in detecting cyberbullies. Social network features, although they do not contribute to detecting cyberbullies, are found to be significant in detecting cyberbullying dynamics. For elapsed time, although average elapsed time of influence instances is shorter than that of non-influence instances, the experimental results show that it is not significant in detecting the dynamics of cyberbullying.

For future work, we will use additional datasets to reduce the possible effect of sampling bias and consider more different features to improve the overall accuracy of the proposed model. Moreover, we can realize the visualization of cyberbullying dynamics to provide moderators and network administrators a more intuitive view of the dynamics of cyberbullying so that they can take measures to control the spread of cyberbullying. For instance, we can build an interface to show how cyberbullying behavior spread over the network in real time and it can help moderators or network administrators conveniently monitor take measures to prevent the spread of cyberbullying. The basic idea of user interface design is as follows: when a thread id is input, the graph of user connection within that thread will be displayed in the interface. The number of users in the thread,

the first posting time and the last posting time will also be displayed aside the network graph. Assuming the red node refers to the bully, the yellow node refers to the user that is influenced by the bully and the black node refers to the normal user. Using Figure 6-1 as the example, when the time slider is moved from the left to the right, one of the black nodes becomes the red node, which means the bully appears. As the time slider is continuously moved, some of the black nodes become the yellow nodes, which means these users have become bullies because of the bullying influence.

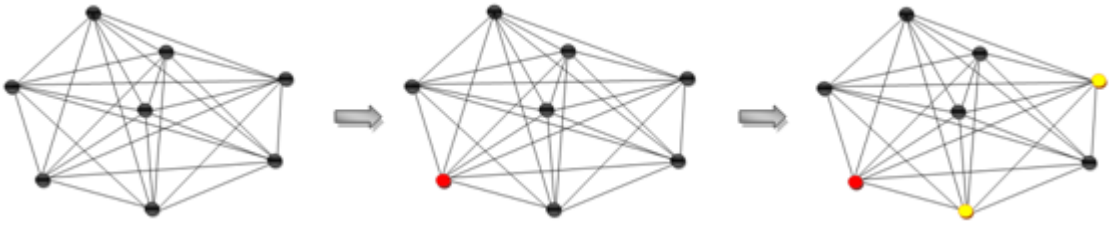


Figure 6-1: Visualization of the Network

When a user node is clicked, the user's information will appear below the visualized graph, including user id, username, gender, age, the number of bullying comments and the corresponding posting time.

References

[1] National Conference of State Legislatures. Cyberbullying. [Online]. Available: http://www.ncsl.org/research/education/cyberbullying.aspx

[2] NoBullying.com. Cyber Bullying Statistics 2014. [Online]. Available: http://nobullying.com/cyber-bullying-statistics-2014/

[3] Ditch the Label. The Annual Cyberbullying Survey in 2013. [Online]. Available: http://www.ditchthelabel.org/the-cyber-bullying-survey-2013/

[4] R. M. Kowalski, S. P. Limber, S. Limber and P. W. Agatston. (2008). Cyberbullying: Bullying in the digital age. *Wiley-Blackwell*.

[5] S. Hinduja and J. W. Patchin. (2009). Bullying Beyond the Schoolyard: Preventing and Responding to Cyberbullying. *Corwin Press*.

[6] S. Hinduja and J. W. Patchin. (2007). Offline Consequences of Online Victimization: School Violence and Delinquency. *Journal of School Violence*, 6(3), 89–112.

[7] S. Hinduja and J. W. Patchin. (2010). Bullying, Cyberbullying, and Suicide. *Archives of Suicide Research*, 14(3), 206–221.

[8] A. Bloxham. (2008). Teenager Sam Leeson Hanged Himself Over 'Emo' Taunts. *The Telegraph*. [Online]. Available: http://www.telegraph.co.uk/news/uknews/2176009/Teenager-Sam-Leeson-hanged-himself-over-Emo-taunts.html

[9] M. L. Ybarra and K. J. Mitchell. (2004). Youth Engaging in Online Harassment: Associations with Caregiver–Child Relationships, Internet use, and Personal Characteristics. *Journal of Adolescence,* 27, 319–336.

[10] S. M. Kift, M. A. Campbell and D. A. Butler. (2010). Cyberbullying in Social Networking Sites and Blogs: Legal Issues for Young People and Schools. *Journal of Law, Information and Science,* 20(2), pp. 60-97.

[11] C. L. Betz. (2011). Cyberbullying: The Virtual Threat. *Journal of Pediatric Nursing*, Volume 26, Issue 4, pp. 283-284.

[12] F. Mishna, M. K. Kassabri, T. Gadalla and J. Daciuk. (2011). Risk Factor for Involvement in Cyber Bullying: Victims, Bullies and Bully-Victims. *Children and Youth Services Review*, 34 (2012), 63–70.

[13] R. A. Sabella, J. W. Patchin and S. Hiduja. (2013). Cyberbullying Myths and Realities. *Computers in Human Behavior*, 29 (2013), 2703–2711.

[14] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis and L. Edwards. (2009). Detection of Harassment on Web 2.0. *CAW2.0 2009,* April 21, 2009, Madrid, Spain.

[15] J. Bayzick, A. Kontostathis and L. Edwards. (2011). Detecting the Presence of Cyberbullying Using Computer Software. *WebSci '11*, June 14-17, 2011, Koblenz, Germany.

[16] K. Dinakar, R. Reichart and H. Liberman. (2011). Modeling the Detection of Textual Cyberbullying. *Association for the Advancement of Artificial Intelligence*.

[17] K. Reynolds, A. Kontostathis and L. Edwards. (2011). Using Machine Learning to Detect Cyberbullying. *The 10th International Conference on Machine Learning and Applications Workshops*, December 2011. Honolulu, HI.

[18] Y. Chen, Y. Zhou, S. Zhu and H. Xu. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. *IEEE International Conference on Social Computing*, Amsterdam, Netherlands.

[19] A. Kontostathis, K. Reynolds, A. Garron and L. Edwards. (2013). Detecting Cyberbullying: Query Terms and Techniques. *WebSci'13*, May 2–4, 2013, Paris, France.

[20] H. Liberman, K. Dinakar and B. Jones. (2011). Let's Gang Up On Cyberbullying. *IEEE Computer Society*, September 2011.

[21] P. Galan-Garcia, J. G. de. La. Puerta, C. L. Gomez, I. Santos and P. G. Bringas. (2013). Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying. *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*, Volume 239, 2014, pp 419-428.

[22] M. Dadvar, D. Trieschnigg, R. Ordelman and F. de Jong. (2013). Improving Cyberbullying Detection with User Context. *Advances in Information Retrieval*, Volume 7814, 2013, pp 693-696.

[23] Q. Huang, V. K. Singh and P. K. Atrey. (2014). Cyber Bullying Detection Using Social and Textual Analysis. *SAM'14*, November 7, 2014, Orlando, Florida, USA.

[24] M. Dadvar, D. Trieschnigg and F. de Jong. (2013). Expert Knowledge for Automatic Detection of Bullies in Social Networks. *The 25th Benelux Conference on Artificial Intelligence*, November 7-8 2013, Delft, Netherlands.

[25] M. Dadvar, D. Trieschnigg and F. de Jong. (2014). Experts and Machines Against Bullies: A Hybrid Approach to Detect Cyberbullies. *Advances in Artificial Intelligence*, Volume 8436, 275-281.

[26] PureSight. [Online]. Available: http://www.puresight.com/

[27] Stopit. [Online]. Available: http://stopitcyberbully.com/

[28] BRIM. [Online]. Available: https://antibullyingsoftware.com/

[29] Wikipedia. Decision Tree Learning. [Online]. Available: http://en.wikipedia.org/wiki/Decision_tree_learning

[30] J. R. Quinlan. (1986). Induction of Decision Trees. *MACH. LEARN*, 1, 81-106.

[31] Wikipedia. ID3 Algorithm. [Online]. Available: http://en.wikipedia.org/wiki/ID3_algorithm

[32] J. R. Quinlan. (1993). C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers, Inc.*

[33] Wikipedia. C4.5 Algorithm. [Online]. Available: http://en.wikipedia.org/wiki/C4.5_algorithm

[34] J. R. Quinlan. (1996). Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence Research* 4 (1996), pp 77-90.

[35] Wikipedia. Information Gain Ratio. [Online]. Available:

http://en.wikipedia.org/wiki/Information_gain_ratio

[36] Wikipedia. Sensitivity and Specificity. [Online]. Available:

http://en.wikipedia.org/wiki/Sensitivity_and_specificity

[37] A. Kontostathis. Data Download. [Online]. Available:

http://www.chatcoder.com/DataDownload

[38] L. von Ahn. Offensive/Profane Word List. [Online]. Available:

http://www.cs.cmu.edu/~biglou/resources/bad-words.txt

[39] Semantria. [Online]. Available: https://semantria.com/

[40] T. Opsahl, F. Agneessens and J. Skvoretz. (2010). Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths. *Social Networks*, April 20, 2010.

[41] K. Okamoto, W. Chen and X. Li. (2008). Ranking of Closeness Centrality for Large-Scale Social Networks. *Frontiers in Algorithmics*, Volume 5059, 186-195.

[42] M. E. J. Newman. (2004). Analysis of Weighted Networks. *Phys. Rev*. E 70, 056131, 24 November 2004.

[43] A. Barrat, M. Barthélemy, R. Pastor-Satorras and A. Vespignani. (2004). The Architecture of Complex Weighted Networks. *The National Academy of Sciences of the USA*, 101(11): 3747–3752.

[44] Wikipedia. Naïve Bayes Classifier. [Online]. Available:

https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[45] Wikipedia. Support Vector Machine. [Online]. Available:

https://en.wikipedia.org/wiki/Support_vector_machine

[46] WEKA. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/