JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Procedural Design of Preventing the Cyber bullying Using Machine Learning Techniques

KSS Praveen Kumar M Kalpana Devi C Narasimham

Praveen.kanakala@gmail.com dkalpananaik@gmail.com drchallan@gmail.com

Abstract

Now a Days ONLINE social networks (OSNs) are frequently flooded with scathing remarks against individuals or organizations on their perceived wrongdoing. Public shaming in online social networks and related online public forums like Twitter has been increasing in recent years. These events are known to have a devastating impact on the victim's social, political, teenagers and financial life. As this is becoming more popular still it faces some common limitations in terms of controlling the abused or vulgar conversations from individual user account. In general, cyber bullying is defined as the process of threatening a small kid or preteen by another child using the internet with some bad conversations and tries to make them feel nervous with these conversations. This is becoming a very severe problem in the current social media by afflicting the children, young adults with these rude messages. Hence, we try to construct a machine learning approach for detecting the set of abused words in the conversations and try to block such conversations not to be spread to others. By using this proposed approach, we can able to create a positive and safe communication in social media. Here we constructed a filter using Support Vector Machine (SVM) to filter out the abused words during communication.

Introduction

SOCIAL Media, as defined is a group of Internet based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content. Via social media, people can enjoy enormous information, convenient communication experience and so on. However, social media may have some side effects such as cyber bullying, which may have negative impacts on the life of people, especially children and teenagers. For bullies, they are free to hurt their peers' feelings because they do not need to face someone and can hide behind the Internet. For victims, they are easily exposed to harassment since all of us, especially youth, are constantly connected to Internet or social media. As reported in [2], cyber bullying victimization rate ranges from 10% to 40%. In the United States, approximately 43% of teenagers were ever bullied on social media [3]. The same as traditional bullying, cyber bullying has negative, insidious and sweeping impacts on children [4], [5], [6]. The outcomes for victims under cyber bullying may even be tragic such as the occurrence of self-injurious behavior or suicides. One way to address the cyber bullying problem is to automatically detect and promptly report bullying messages so that proper measures can be taken to prevent possible tragedies. Previous works on computational studies of bullying have shown that natural language processing and machine learning are powerful tools to study bullying [7], [8]. Cyber bullying detection can be formulated as a supervised learning problem.

In cyber bullying detection, the numerical representation for Internet messages should be robust and discriminative. Since messages on social media are often very short and contain a lot of informal language and misspellings, robust representations for these messages are required to reduce their ambiguity. Even worse, the lack of sufficient high-quality training data, i.e., data sparsity make the issue more challenging. Firstly, labeling data is labor intensive and time consuming. Secondly, cyber bullying is hard to describe and judge from a third view due to its intrinsic ambiguities. Thirdly, due to protection of Internet users and privacy issues, only a

small portion of messages are left on the Internet, and most bullying posts are deleted. As a result, the trained classifier may not generalize well on testing messages that contain non activated but discriminative features.

In addition, the performance of these approaches rely on the quality of hand-crafted features, which require extensive domain knowledge. In this paper, we investigate one deep learning method named stacked denoising auto encoder (SDA) [15]. SDA stacks several denoising auto encoders and concatenates the output of each layer as the learned representation. Each denoising auto encoder in SDA is trained to recover the input data from a corrupted version of it. The input is corrupted by randomly setting some of the input to zero, which is called dropout noise. This denoising process helps the auto encoders to learn robust representation. In addition, each auto encoder layer is intended to learn an increasingly abstract representation of the input [16]. In this paper, we develop a new text representation model based on a variant of SDA: marginalized stacked denoising auto encoders (mSDA) [17], which adopts linear instead of nonlinear projection to accelerate training and marginalizes infinite noise distribution in order to learn more robust representations. We utilize semantic information to expand mSDA and develop Semantic-enhanced Marginalized Stacked Denoising Auto encoders (smSDA). The semantic information consists of bullying words. An automatic extraction of bullying words based on word embeddings is proposed so that the involved human labor can be reduced. During training of smSDA, we attempt to reconstruct bullying features from other normal words by discovering the latent structure, i.e. correlation, between bullying and normal words. The intuition behind this idea is that some bullying messages do not contain bullying words. The correlation information discovered by smSDA helps to reconstruct bullying features from normal words, and this in turn facilitates detection of bullying messages without containing bullying words.

For example, there is a strong correlation between bullying word fuck and normal word off since they often occur together. If bullying messages do not contain such obvious bullying features, such as fuck is often misspelled as fck, the correlation may help to reconstruct the bullying features from normal ones so that the bullying message can be detected. It should be noted that introducing dropout noise has the effects of enlarging the size of the dataset, including training data size, which helps alleviate the data sparsity problem. In addition, L1 regularization of the projection matrix is added to the objective function of each autoencoder layer in our model to enforce the sparsity of projection matrix, and this in turn facilitates the discovery of the most relevant terms for reconstructing bullying terms. The main contributions of our work can be summarized as follows: Our proposed Semantic-enhanced Marginalized Stacked Denoising Autoencoder is able to learn robust features from BoW Representation in an efficient and effective way. These robust features are learned by reconstructing original input from corrupted (i.e., missing) ones. The new feature space can improve the performance of cyberbullying detection even with a small labeled training corpus. Semantic information is incorporated into the reconstruction process via the designing of semantic dropout noises and imposing sparsity constraints on mapping matrix. In our framework, high-quality semantic information, i.e., bullying words, can be extracted automatically through word embedding Finally, these specialized modifications make the new feature space more discriminative and this in turn facilitates bullying detection. Comprehensive experiments on real-data sets have verified the performance of our proposed model.

Clustering is refereed together of the foremost process in processing which is employed for separating a group of un-supervised data into a meaningful way. This process could even be how in process of data discovery which successively uses clustering mechanism for getting data accurately. Clustering algorithms are typically used for exploratory data analysis, where there's little or no prior knowledge about the info [1], [2]. This is most often used in several applications of computer data inspection, including the one addressed in our work. If we come across the technical viewpoint, the input data is initially contained several objects, where some are not labeled and may be found are a priori unknown. Moreover, even if we try to assume the labeled datasets might be available from previous analyses, there is no complete hope and assurance of getting a valid outcome after a deep investigation process.

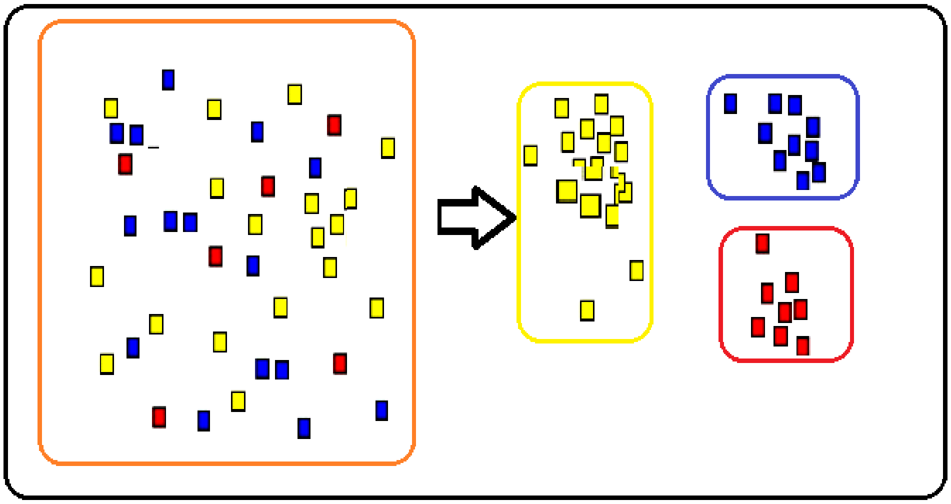


Figure. 1 Represent the Example of Data Clustering Technique In The Process Of Data Mining

From the above figure 1, we will clearly see the detailed and clear example for clustering technique within the process of knowledge mining. Initially, we attempt to collect the un-supervised data as input which aren't arranged so as then we attempt to apply any clustering algorithm and mine the info during a clustered manner. Here we attempt to assume three different blocks with different colours. Here initially all the blocks are kept in an unsupervised manner and now we attempt to apply the clustering algorithm so as to separate the things supported some input function. Here the info which is unsupervised is nearly some colour blocks which are randomly shuffled into one group and where each and each colour block has individual characteristics in appearance and shape. Now we attempt to apply the clustering algorithm K Means so as to categorize the color blocks into separate groups. Now the color blocks which are having an equivalent colour inherit one block and that they are termed together cluster and people which are having a special appearance as treated as separate blocks and that they are treated as unstructured data which remains not matched with any of those groups. During this same way, we will apply an equivalent clustering algorithm on all examples to cluster the info into various individual groups.

Cyberbullying are often defined as aggressive, intentional actions performed by a private or a gaggle of individuals via data communication methods like sending messages and posting comments against a victim. Different from traditional bullying that sometimes occurs at college during face-to-face communication, cyberbullying on social media can happen anywhere at any time. For bullies, they're liberal to hurt their peers' feelings because they are doing not got to face someone and may hide behind the web. For victims, they're easily exposed to harassment since all folks, especially youth, are constantly connected to the web or social media. Bag-of-words (BoW) model is one commonly used model that every dimension corresponds to a term. By mapping all abused words into fixed-length vectors, the learned representation are often further processed for varied language processing tasks. Here the Support vector machine is used for gathering all abused distinct words and place all those words in that BoW model. Once the words are collected and placed in the BoW database and it is clearly shown in figure 2, then the messages are compared with that bag of words and if any message contains a word found from Bow, then they are identified as cyberbullying conversation and the system should automatically block all such abused conversations.

Cyberbullying Detection Algorithms on Social Media:

In this work uses “Bag of Words”, “The Porter stemming algorithm”, “Support Vector Machine Algorithm”.

1. I have chosen 5 types of labels in which the sentences are categorized . They are HATE, VULGAR, OFFENSE, SEX, and VIOLENCE.
2. We add the above-mentioned labels in the data sets with words.
3. User exchange the comments on SNM(Social Network Media) using Porter Stemming Algorithm and SVM. We divide entire comment into words and identify the noun and verbs along with also identify how many labelled words are used in the comments.
4. Admin will block the comments from review if more label words are used else caution will be sent.

Algorithm: BOW (Bag of words)

A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:

1. A vocabulary of known words.
2. A measure of the presence of known words.

Step 1: Collect Data

- Convert text to lower case.
- Remove all non-word characters.
- Remove all punctuations.

Step 2: Design the Vocabulary

Now we can make a list of all of the words in our model vocabulary.

The unique words here (ignoring case and punctuation) are:

Step 3: Create Document Vectors

Building the Bag of Words model

In this step we construct a vector, which would tell us whether a word in each sentence is a frequent word or not. If a word in a sentence is a frequent word, we set it as 1, else we set it as 0.

The next step is to score the words in each document.

The simplest scoring method is to mark the presence of words as a boolean value, 0 for absent, 1 for present.

There are simple text cleaning techniques that can be used as a first step, such as:

- Ignoring case
- Ignoring punctuation
- Ignoring frequent words that don't contain much information, called stop words, like "a," "of," etc.
- Fixing misspelled words.
- Reducing words to their stem (e.g. "play" from "playing") using stemming algorithms.

A more sophisticated approach is to create a vocabulary of grouped words. This both changes the scope of the vocabulary and allows the bag-of-words to capture a little bit more meaning from the document. In this approach, each word or token is called a "gram". Creating a vocabulary of two-word pairs is, in turn, called a bigram model. Again, only the bigrams that appear in the corpus are modelled, not all possible bigrams. An N-gram is an N-token sequence of words: a 2-gram (more commonly called a bigram) is a two-word sequence of words like "please turn", "turn your", or "your homework", and a 3-gram (more commonly called a trigram) is a three-word sequence of words like "please turn your", or "turn your homework".

The Porter stemming algorithm:

The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and in flexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems.

Support Vector Machine Algorithm

By using Support Vector Machine, we can identify the good and bad words graphically and also filter them. It is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, perform classification by finding the hyper-plane that differentiates the two classes very well .

Existing system and Limitations:

In the existing system there was no pre-defined method or software to classify the abused or cyber bulling messages for a text message which is posted on OSN walls and identify the meaning of that word and block that message not to be posted directly on the users wall. So the following are the limitations that take place in the existing system. They are as follows:

1. Till now there was no method in OSN literature. (to automatically detect the cyber bullying messages and encode them into a separate list.)
2. There was no classification algorithm in literature .(that can automatically read all the text which is posted by the users and recognize if there are any abused content available on that posted messages.)
3. There is no term like **BoW** in the existing system, where a bag of words is listed into a database and these bag of words are used for matching the dimensions of corresponding term which is posted on the wall.
4. The main limitation of BoW is this can identify the exact word in exact message if the same message contains the word in plural way, this can’t be identified as matched word.
5. In the existing there is no concept like segregate the messages into categories like cyber bulled messages and Non-Cyber Bulled category message.

Proposed system

In the proposed system used expert knowledge for feature learning. The proposed system use ML-Approach for classifying the semantic meanings of posted message and try to combine BoW features, sentiment features and contextual features to train a support vector machine for online harassment detection. Here in our proposed system as an extension and also designed a label specific features to extend the general features, where the label specific features are learned by Linear Discriminative Analysis**.** Here by using this label specific feature and can able to get the count of abused or harassed words that are repeated and used within the posted message. Hence we try to construct a machine learning approach for detecting the set of abused words in the conversations and try to block such conversations not to be spread to others. By using this proposed approach we can able to create a positive and safe communication in social media. Here we constructed a filter using Support Vector Machine (SVM) to filter out the abused words during communication.

ADVANTAGES OF PROPOSED SYSTEM

1. Now there was method like SVM classification to automatically detect the cyber bullying messages and encode them into a separate list. By using this can avoid the problem.
2. By using cyberbullying now can automatically read all the text which is posted by the users and recognize if there are any abused content available on that posted messages.
3. Consider the BoW system, by using this can maintain a bag of words is listed into a database and these bag of words are used for matching the dimensions of corresponding term which is posted on the wall.
4. The advantage of cyberbullying to segregate the messages into categories like cyber bulled messages and Non-Cyber Bulled category message.

Implementation is the stage where the theoretical design is converted into programmatically manner. In this stage we will divide the application into a number of modules and then coded for deployment. The front end of the application takes JSP, HTML and Java Beans and as a Back-End Data base we took My SQL data base. The application is divided mainly into following 4 modules. They are as follows:

1. Network Creation Module
2. Construction of Bullying Feature Set Module
3. Label Feature Selection Module
4. Identifying Cyberbullied Users

NETWORK CREATION MODULE

In this module initially we need to construct a network containing single admin and multiple users. Where the admin has the facility to add a set of words into each BoW database based on individual category. The admin should add each and every word into the database individually. Once if a word is added in one category the same word shouldn't be added on another category. So, this should be mandatory step for the admin while adding words into the database. Also, admin has the facility to authorize each and every user at the time of registration. The user who got activated by admin only can access his profile by login into the site. Those users who are not authorized can't be entering into their individual accounts at any cost.

CONSTRUCTION OF BULLYING FEATURE SET MODULE

The bullying features play an important role and should be chosen properly. In the following, the steps for constructing bullying feature set Z_b are given, in which the first layer and the other layers are addressed separately. Here we try to add all bullying words based on category wise and try to maintain a vector to hold all these data.

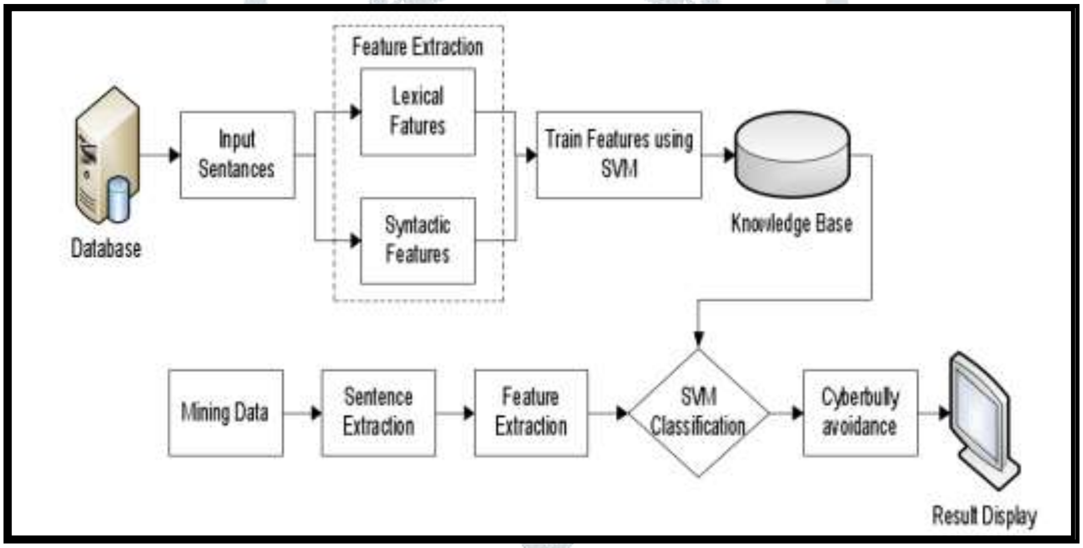
LABEL FEATURE SELECTION MODULE

Here we proposed a labeled Feature Selection method where the labeling is done because, if any word is matched from a set of Bow, then they are automatically identified as an abused word and they will be identified based on individual category wise. Hence labeled based feature selection method is mainly used for categorizing each and every matched word based on category wise. Here we try to use SVM algorithm which maintain BoW database with set of labels like Sex, Vulgar, Offensive, Hate and Violence.

IDENTIFYING CYBERBULLED USERS

In this module we try to create a separate list and try to classify the users who try to post normal messages under one category and those who try to post abuse conversations in separate list.Here the users who try to post a message either comment or reply by using any bullying words is automatically identified by the admin and they are tagged as Cyberbulled user and those details can be monitored by the admin.

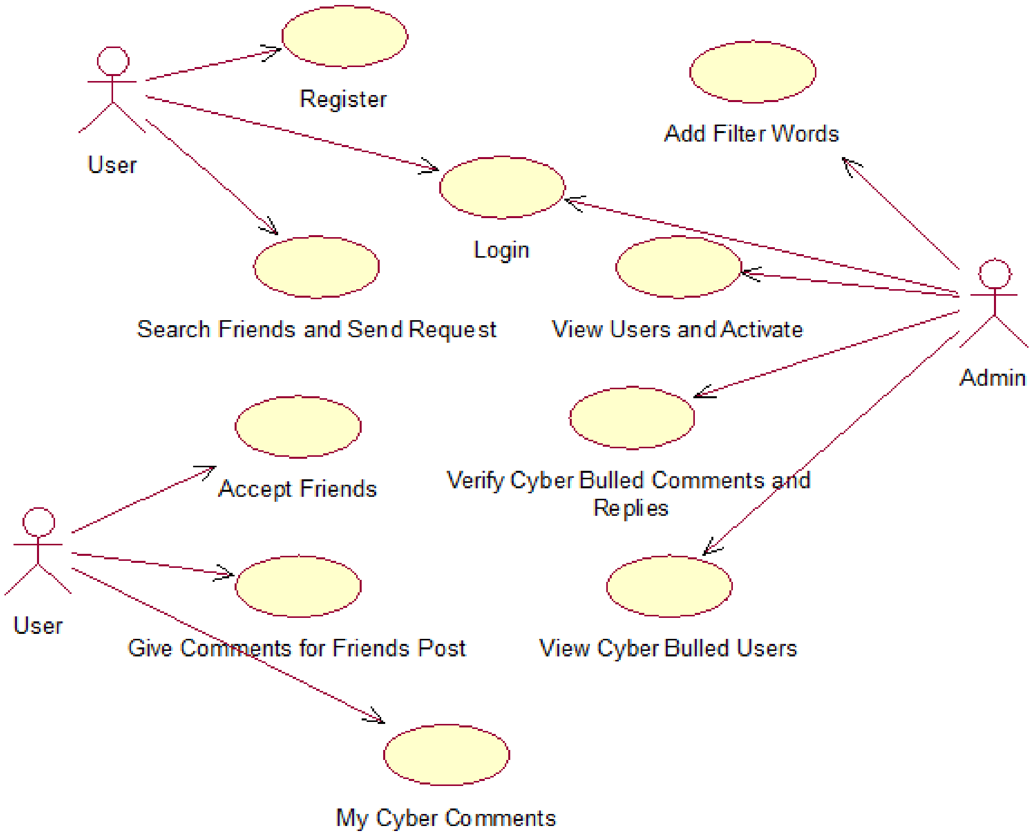
System Architecture:



System Architecture

USE CASES

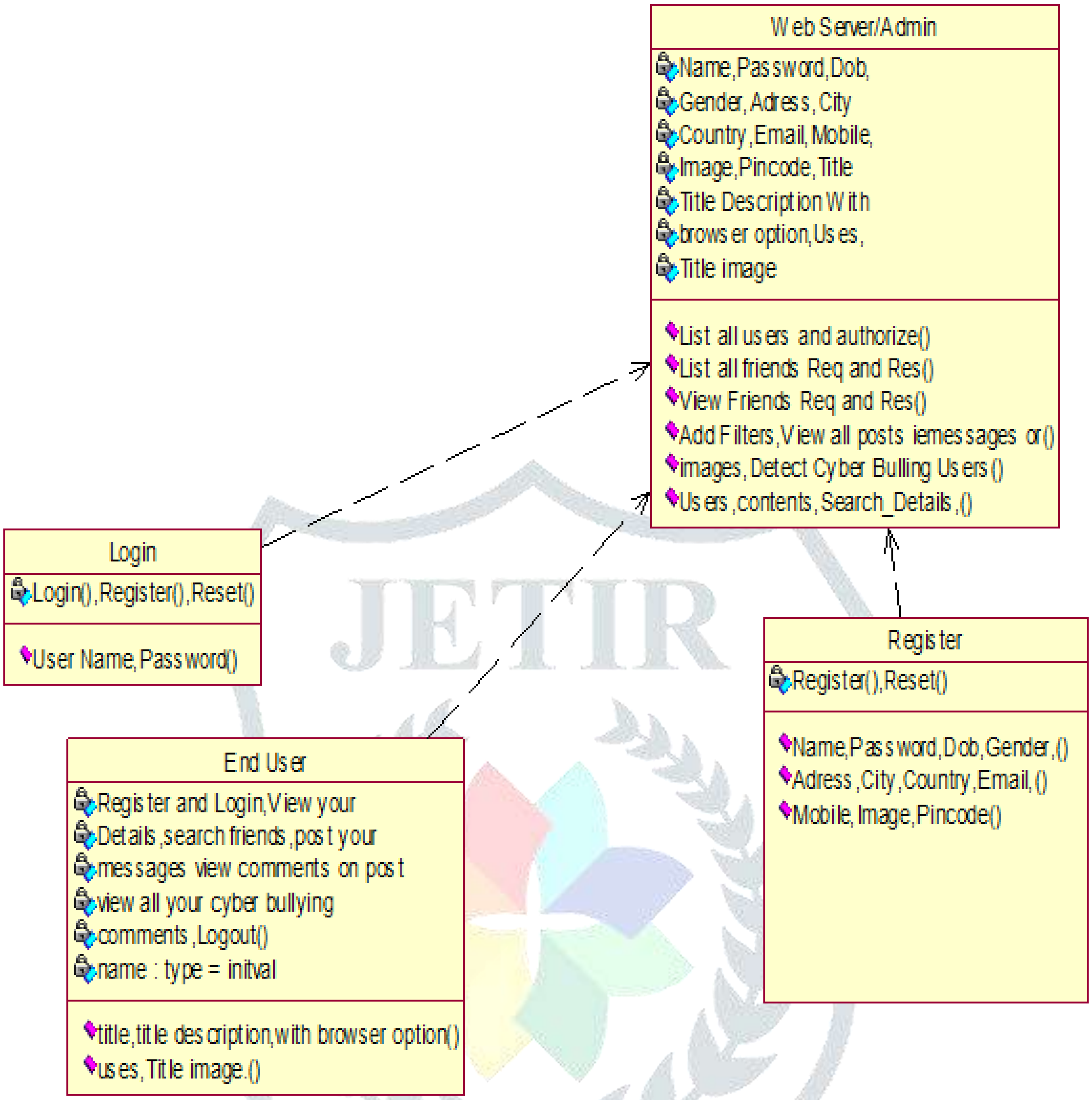
A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



Use Case Diagram

CLASS DIAGRAM

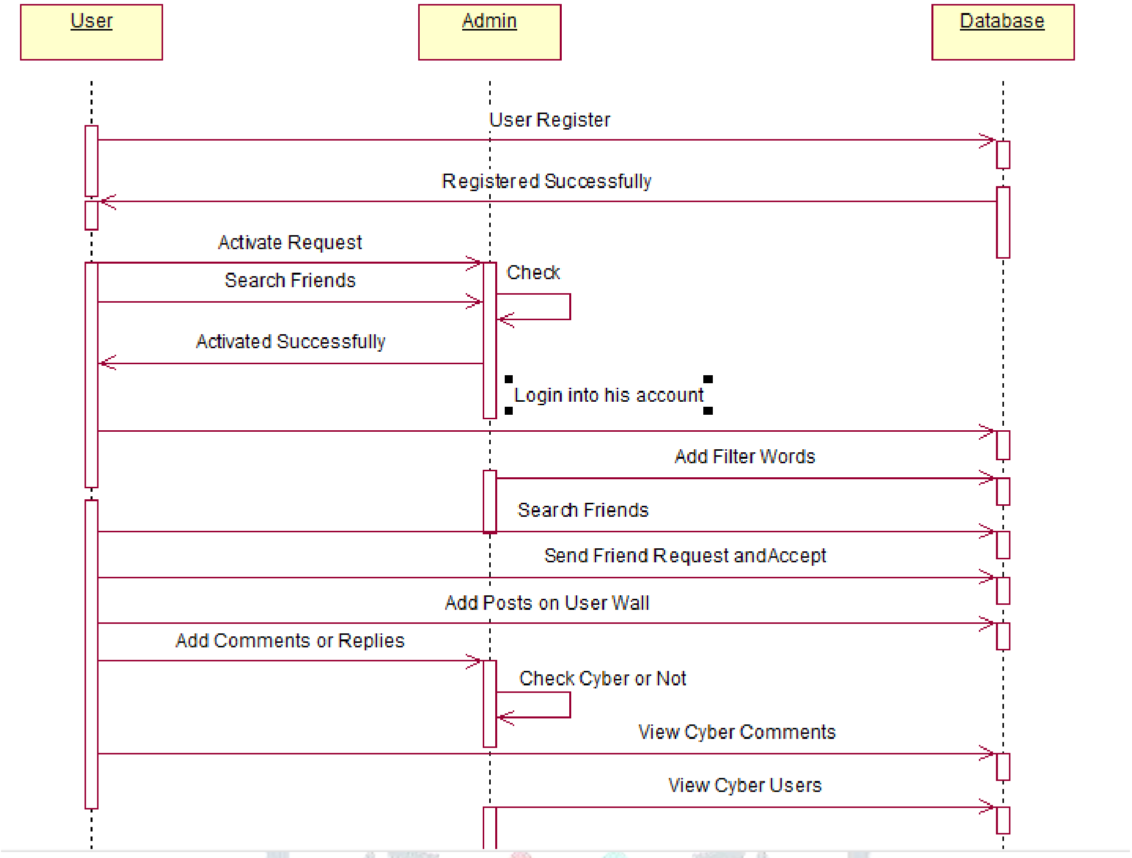
In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes.


Web Server/Admin
Name,Password,Dob,
Gender,Adress,City
Country,Email,Mobile,
Image,Pincode,Title
Title Description With
browser option,Uses,
Title image
List all users and authorize()
List all friends Req and Res()
View Friends Req and Res()
Add Filters,View all posts iemessages or()
images,Detect Cyber Bulling Users()
Users,contents,Search_Details,()
Login
Login(),Register(),Reset()
User Name,Password()
Register
Register(),Reset()
Name,Password,Dob,Gender,()
Adress,City,Country,Email,()
Mobile,Image,Pincode()
End User
Register and Login,View your
Details,search friends,post your
messages view comments on post
view all your cyber bullying
comments,Logout()
name : type = initval
title,title description,with browser option()
uses,Title image.()

Class Diagram

Sequence Diagram:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

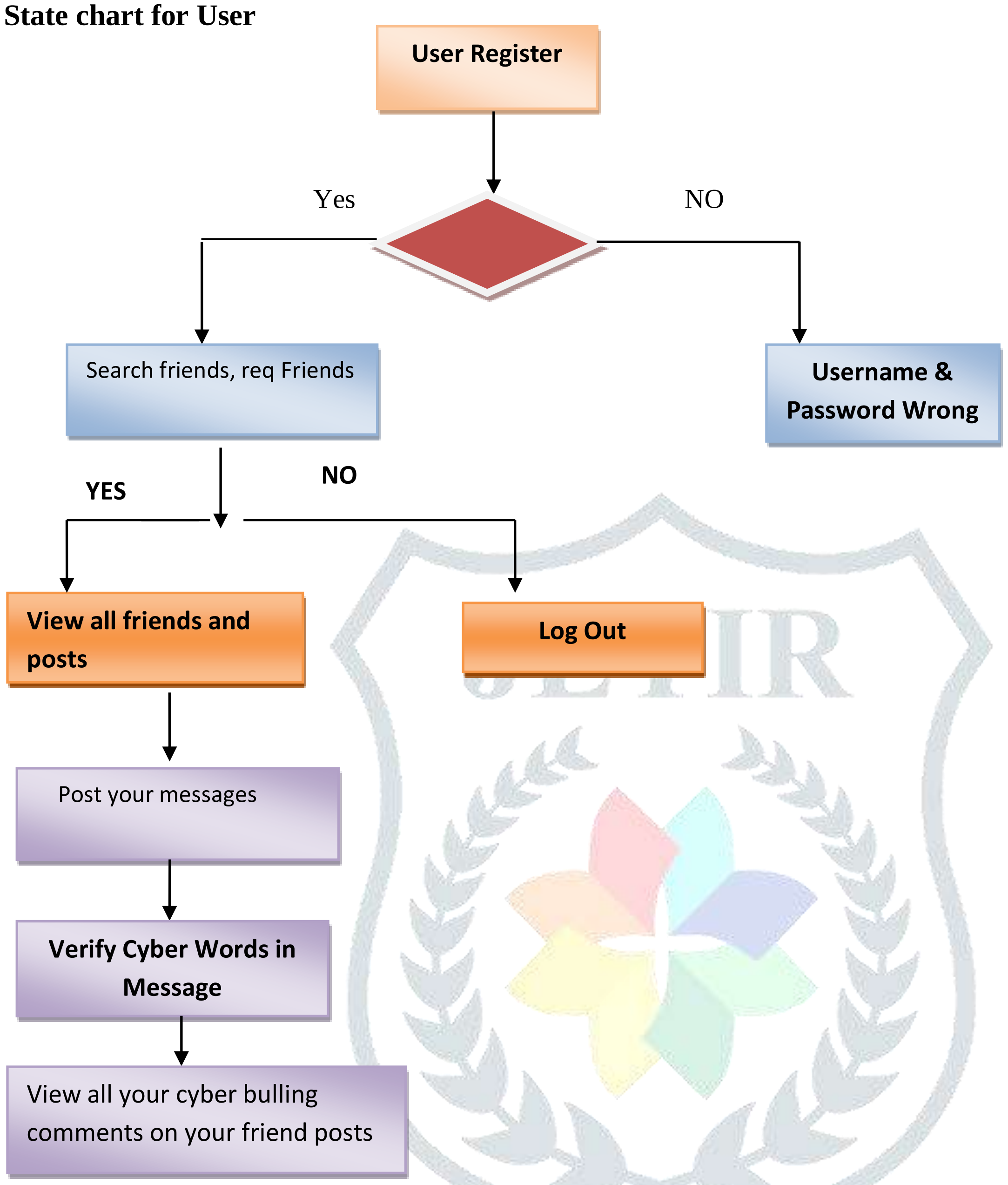


Sequence Diagram

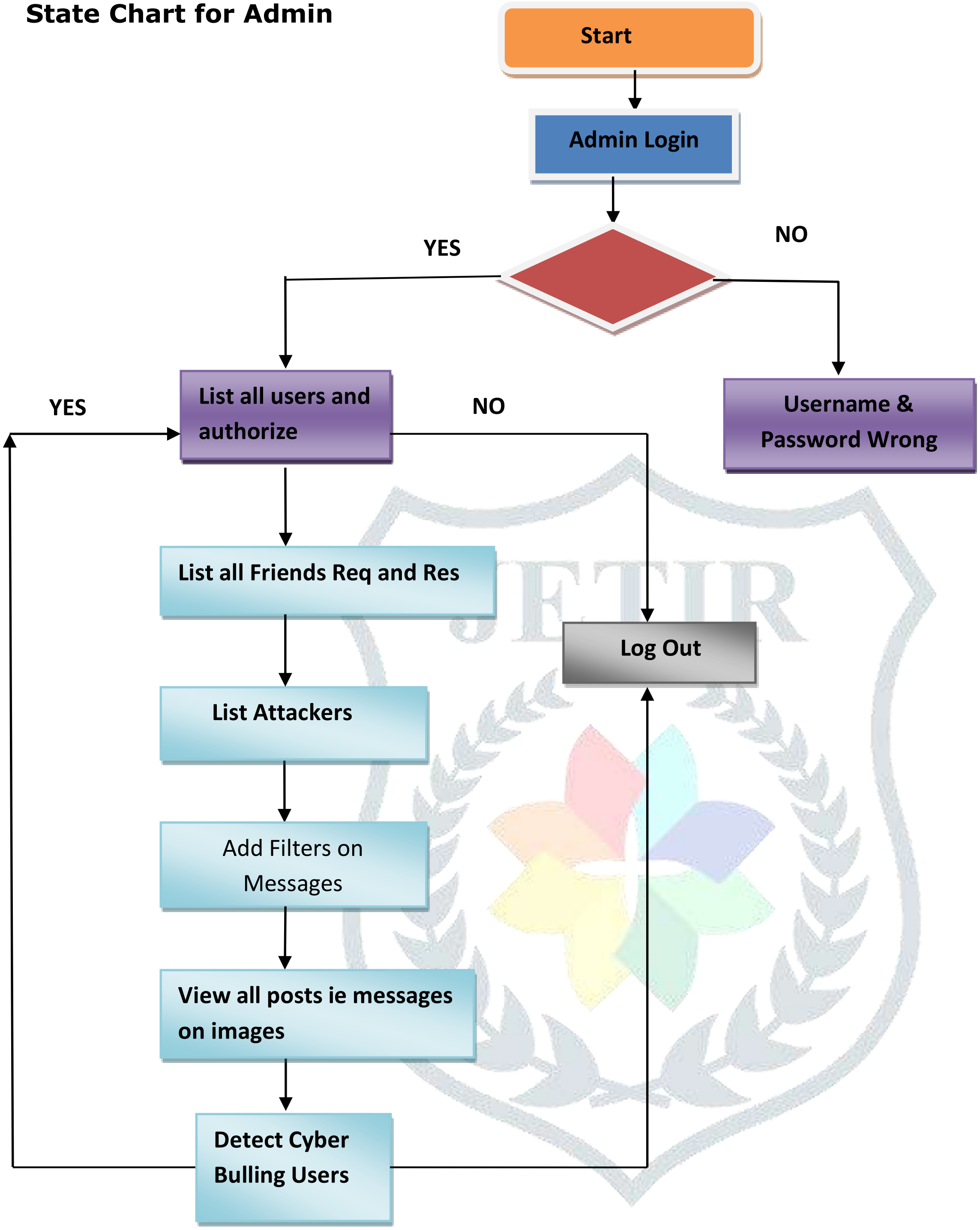
State chart Diagram

State chart diagram describes the flow of control from one state to another state. States are defined as a condition in which an object exists and it changes when some event is triggered. The most important purpose of State chart diagram is to model lifetime of an object from creation to termination. State chart diagrams are also used for forward and reverse engineering of a system. However, the main purpose is to model the reactive system.

State chart for User


User Register
Yes
NO
Search friends, req Friends
Username & Password Wrong
YES
NO
View all friends and posts
Log Out
Post your messages
Verify Cyber Words in Message
View all your cyber bulling comments on your friend posts

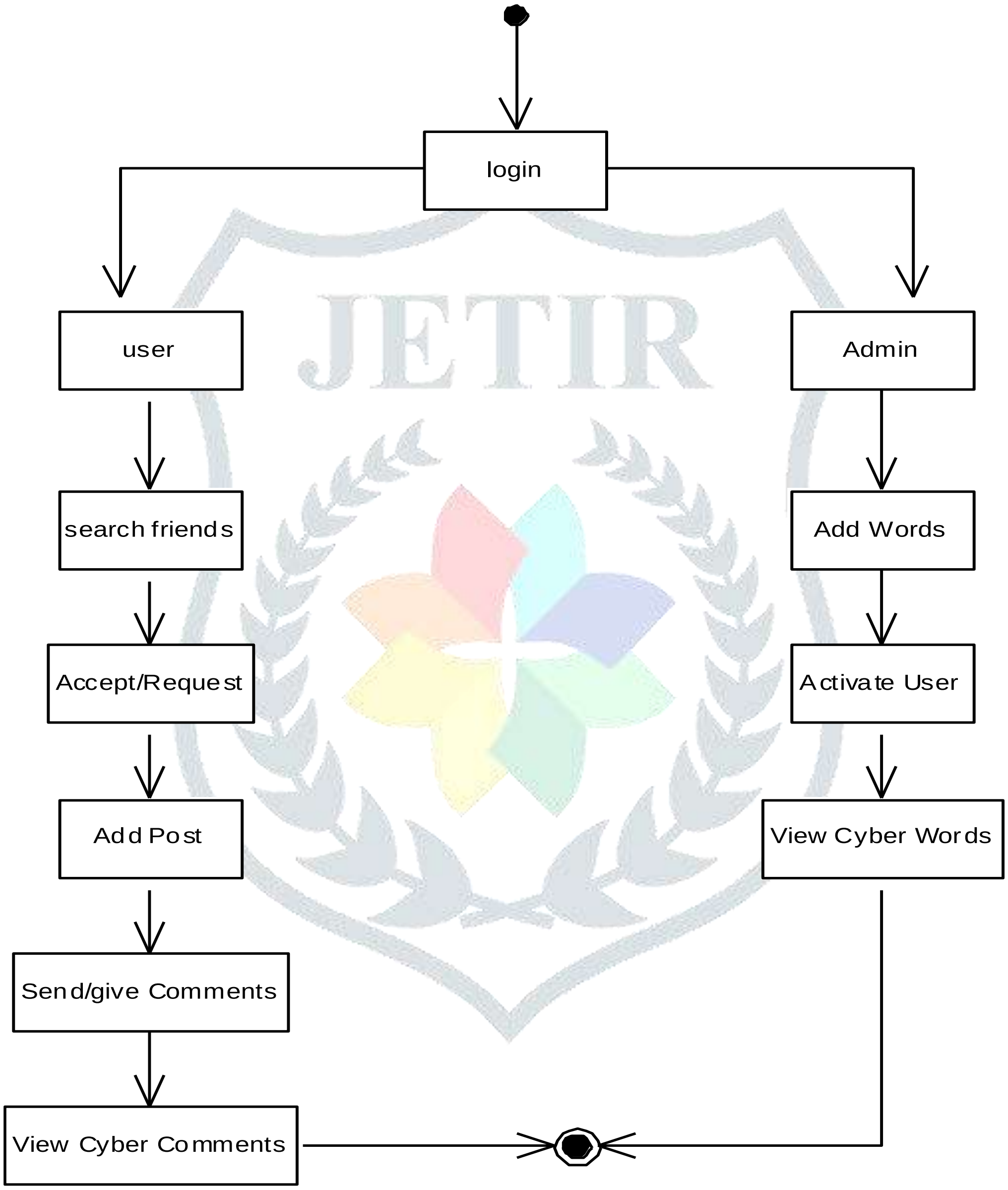
State Chart for Admin


Start
Admin Login
YES
NO
List all users and authorize
YES
NO
Username & Password Wrong
List all Friends Req and Res
Log Out
List Attackers
Add Filters on Messages
View all posts ie messages on images
Detect Cyber Bulling Users

State Chart Diagram for Admin

Activity diagram

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



Activity Diagram

Conclusions and Future scope

From the proposed idea, it is clearly known 60-70% of text cyberbullying can be avoided from posting and also reporting the person helps us to improve that people does not miss something they want to know. This helps in preventing a person from getting bullied and also protects the internet users from cyberbullying crimes. Using sentiment analysis polarity of the text has been defined and also analyzing the text from corpus helps in identifying the most used cyberbullying text. In addition, word embedding has been used to automatically expand and refine bullying words list that is initialized by domain knowledge. The performance of our approaches has been experimentally verified through two cyberbullying corpora from social Medias: Twitter and MySpace. It can be further improve the robustness of the learned representation by considering word order in messages and also using natural language processing techniques in order to predict any cyberbullying words which are not in the dataset and add the same by means of feedback into BoW database.

References:

[1] A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of social media,” Business horizons, vol. 53, no. 1, pp. 59–68, 2010.

[2] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, “Bullying in the digital age: A critical review and metaanalysis of cyberbullying research among youth.” 2014.

[3] M. Ybarra, “Trends in technology-based sexual and non-sexual aggression over time and linkages to nontechnology aggression,” National Summit on Interpersonal Violence and Abuse Across the Lifespan: Forging a Shared Agenda, 2010.

[4] B. K. Biggs, J. M. Nelson, and M. L. Sampilo, “Peer relations in the anxiety–depression link: Test of a mediation model,” Anxiety, Stress, & Coping, vol. 23, no. 4, pp. 431–447, 2010.

[5] S. R. Jimerson, S. M. Swearer, and D. L. Espelage, Handbook of bullying in schools: An international perspective. Routledge/Taylor & Francis Group, 2010.

[6] G. Gini and T. Pozzoli, “Association between bullying and psychosomatic problems: A meta-analysis,” Pediatrics, vol. 123, no. 3,pp. 1059–1065, 2009.

[7] A. Kontostathis, L. Edwards, and A. Leatherman, “Text mining and cybercrime,” Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK, 2010.

[8] J.-M. Xu, 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. Association for Computational Linguistics, 2012, pp. 656–666.