Automated Detection of Cyberbullying Using Machine Learning

Niraj Nirmal[1], Pranil Sable[2], Prathamesh Patil[3], Prof. Satish Kuchiwale[4]

[1-4]*SIGCE Maharashtra, INDIA.*

Abstract— *Increasing the use of Internet and facilitating access to online communities such as social media have led to the emergence of cybercrime. Cyberbullying is very common now a days. which have no tracking like it may harm any individual, business, society, country in past few days it seems that riots were happened due to some statement used by one community on another its important to identify such content which spreads hate or harm community text processing, NLP (natural language processing) is an emerging field with the help of NLP and machine learning algorithms such as naive bayes, random forest, SVM we are going to identify cyberbullying in twitter. Objectives of this implementation written in objective section. Image character with the help of OCR will be done by us to find image - based cyberbullying the impact on individual basis thus will be checked on dummy system. Machine learning and natural language processing techniques to identify the characteristics of a cyberbullying exchange and automatically detect cyberbullying by matching textual data to the identified traits. On the basis of our extensive literature review, we categorise existing approaches into 4 main classes, namely supervised learning, lexicon-based, rule-based, and mixed-initiative approaches. Supervised learning-based approaches typically use classifiers such as SVM and Naïve Bayes to develop predictive models for cyberbullying detection.*

Index Terms— cyber bullying, natural language processing, machine learning algorithms, Social networking.

I. Introduction

It is not sufficient to remind students of regulations forbid-ding plagiarism; In recent years, the use of social networking increased. And social networking sites are great tools of connecting to people. However, as social networking has become widespread. People are finding illegal and unethical ways to use these communities. We see that people, especially teens and young adults, are finding new ways to bully one another over the Internet. Bullying is not a new phenomenon and cyber bullying has manifested itself as soon as digital technologies have become primary communication tools. On the positive side, social media like blogs, social networking sites (e.g. Facebook), and instant messaging platforms (e.g. WhatsApp) make it possible to communicate with anyone and at any time. Moreover, they are a place where people engage in social interaction, offering the possibility to establish new relationships and maintain existing friendships. On the negative side however, social media increase the risk of children being confronted with threatening situations including grooming or sexually transgressive behaviour, signals of depression and suicidal thoughts, and cyberbullying. Users are reachable 24/7 and are often able to remain anonymous if desired: this makes social media a convenient way for bullies to target their victims outside the school yard. The detection of cyberbullying and online harassment is often formulated as a classification problem. Techniques typically used for document classification, topic detection, and sentiment analysis can be used to detect electronic bullying using characteristics of messages, senders, and the recipients. It should, however, be noted that cyberbullying detection is intrinsically more difficult than just detecting abusive content. Additional context may be required to prove that an individual abusive message is part of a sequence of online harassment directed at a user for such a message to be labelled as cyberbullying. The growth of cyberbullying activities is increasing as equally as the growth of social networks. Cyberbullying activities poses a significant threat to mental and physical health of the victims. Project about detection of bullying is present but implementation for monitoring social network to detect cyberbullying activities is less. Hence, the proposed system focuses on detecting the presence of cyberbullying activity in social networks using natural language processing and machine learning algorithms which helps government to take action before many users becoming a victim of cyberbullying. Detection of cyberbullying and the provision of subsequent preventive measures are the main courses of action to combat cyberbullying. The proposed method is an effective method to detect cyberbullying activities on social media. The detection method can identify the presence of cyberbullying terms and classify cyberbullying activities in social network such as Flaming, Harassment, Racism and Terrorism using natural language processing and machine learning algorithms. Cyberbullying detection is inherently difficult due to the subjective nature of bullying. It extends beyond detecting negative sentiments or abusive content in a message as these tasks, on their own, do not necessarily mean that the message is in fact bullying. For example, a message such as “I’m disgusted by what you said today and I never want to see you again” is difficult to classify as bullying without understanding the larger context of the exchange, even though the message is clearly expressing very negative sentiments. Conversely, positively-expressed sarcasm.

II. Related Work

Table: Literature Survey

Title	Authors	Problem	Solution	Result
An Effective Approach for Cyberbullying Detection and avoidance	Divyashree, Vinutha H, Deepashree N S	The biggest problem regarding cyberbullying is that the age group of the offenders ranges from as young as eight to the legal adult age of eighteen and beyond. Once happen this activity then victims are often left permanently then difficult to find them.	In this paper focused on the issues of robust system and objectives are 1) Automatic detection and avoidance of cyberbully attack in internet. 2) Effective age authentication for website browsing and categorizing the links based on age. 3) Effective website filtering in search results based on ranking. 4) Enhanced searching procedure promisingly reduces the effort of user in searching indented websites.	represented a novel method on the current scenario of cyber-bullying and various methods available for the detection and prevention of cyber harassment. Our concept depends upon the text analysis, the data which is uploaded or text written by any user is first analyzed.
Using Machine Learning to Detect Cyberbullying	Kelly Reynolds, April Kontostathis, Lynne Edwards	teens and young adults, are finding new ways to bully one another over the Internet. in a study conducted by Symantec reported that, to their knowledge, their child has been involved in a cyberbullying incident.	Used machine learning algorithm to detect cyberbullying. For training the data downloaded from website. The data was labeled using a web service. the labeled data, in conjunction with machine learning techniques provided by the Weka tool kit, to train a computer to recognize bullying content.	used a language-based method of detecting cyberbullying. By recording the percentage of curse and insult words within a post.
Cyberbullying Detection System on Twitter	Liew Choong Hon, Kasturi Dewi Varathan	Increased cyberbullying attacks on the social network services. To prevent these activities proposed an system.	this system, the users can identify the cyberbullying related tweets based on the keywords and populate it in a news feed form. By doing this, it allows users to determine the identities of the cyberbullies and the victims from the cyberbullying tweets	with the advent of this cyberbullying detection and solution system in Twitter, it will help the authorities to monitor, regulate or at least decrease the harassing incidents in cyberspace
Automatic detection of cyberbullying in social media text	Cynthia Van Hee,Gilles Jacobs,Chris Emmery,Bart Desmet,	Increased the cyberbullying using Social media sides/apps.	The focus of this paper is on automatic cyberbullying detection in social media text by modelling posts written by bullies, victims, and bystanders of online bullying. In this paper support vector machine is used to exploiting a rich feature set and investigate which information sources contribute the most for the task.	In this paper investigate the automatic detection of cyberbullying-related posts on social media. Given the information overload on the web, manual monitoring for cyberbullying has become unfeasible. Automatic detection of signals of cyberbullying would enhance moderation and allow to respond quickly when necessary.
Methods for detection of cyberbullying: A survey	Rekha Sugandhi, Anurag Pande, Siddhant Chawla, Abhishek Agrawal, Husen Bhagat	major problem when it comes to cyber bullying is the lack of identifiable parameters which mark any post as a bullying instance.	This paper aims to review the different methods and algorithms used for detection in cyber bullying and provide a comparative study amongst them so as to decide which method is the most effective approach and provides the best accuracy.	In this paper realize support vector machines have given the best result. We plan to implement SVM in our project as the primary classifier for our base dataset.

2.1 Aim of the Project

The main aim of the detecting the cyberbullying model will help to improve manual monitoring for cyberbullying on social networks. In this project we fetch the tweets from twitter accounts and preprocess the twits and images and applying generated model will detect the cyberbullying or not.

The objectives of the systems development and event management are:

Collect the dataset of bullying words and preprocess it and apply natural language processing and then machine learning algorithms Generate different machine learning algorithm model.

Fetch the tweets from twitter account and preprocess it.

Apply generated model on the fetched tweets and get final output cyberbullying or not.

2.2 Scope of the Project

Cyberbullying is the use of electronic communication to bully a person by sending harmful messages using social media, instant messaging or through digital messages. Cyberbullying can be very damaging to adolescents and teens. It can lead to anxiety, depression, and even suicide. Also, once things are circulated on the Internet, they may never disappear, resurfacing at later times to renew the pain of cyberbullying. Cyberbullying can be very damaging to adolescents and teens. It can lead to anxiety, depression, and even suicide. Also, once things are circulated on the Internet, they may never disappear, resurfacing at later times to renew the pain of cyberbullying. So overcome these issues detecting the cyberbullying is very important in now a days which will help to stop cyberbullying on social media networks.

2.3 Problem Statement

The social media network gives us to great communication platform opportunities they also increase the vulnerability of young people to threatening situations online. Cyberbullying on an social media network is a globle phenomenon because of its huge volumes of active users. The trend shows that the cyber bullying on social network is growing rapidly every day. Recent studies report that cyberbullying constitutes a growing problem among youngsters. Successful prevention depends on the adequate detection of potentially harmful messages and the information overload on the Web requires intelligent systems to identify potential risks automatically. So, In this project we focus on to make a model on automatic cyberbullying detection in social media text by modelling posts written by bullies on social network.

III. Methodology

This project we will develop using python and web technology. Within that first we will search and find the the dataset and download it for train the model. After downloading first we will pre-process the data and then transferred to Tf-Idf. Then with the help of naïve bayes, SVM (Support vector machine) and DNN algorithm we train the dataset and generate model separately. Then we are going to develop a web based application using FLASK framework. We will fetch the real time tweets from twitter and then we apply generated model to these fetched tweets and check the text or images are cyberbullying or not. These all-purpose we are using python as backend, Mysql is database and for frontend html, css, javascript etc.

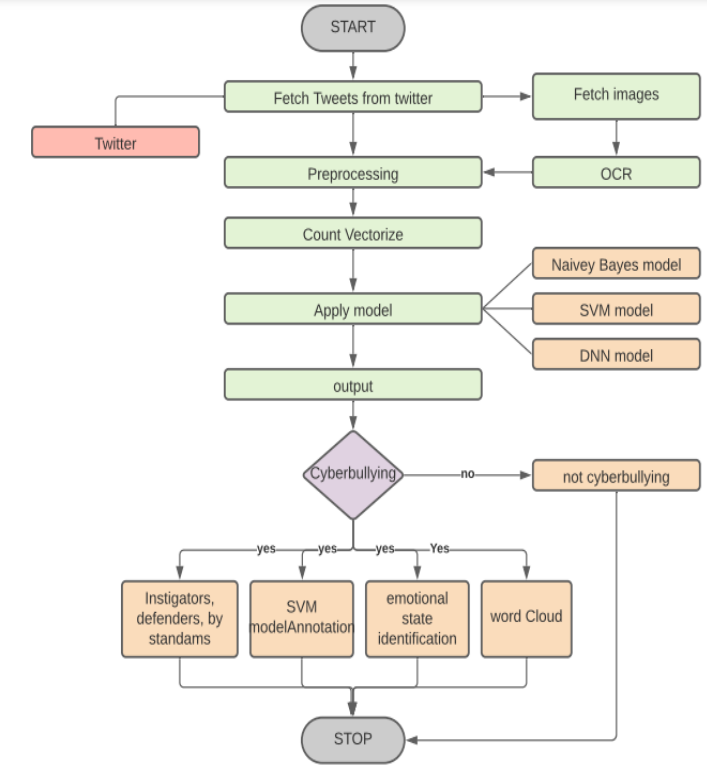

Figure 1: Flow diagram of Cyberbullying Detection

3.1 Technique of detection

3.1.1 Textual Based:

We group features such as cyberbullying keywords, pro - fanity, pronouns, n-grams, Bags-of-words (BoW), Term Frequency Inverse Document Frequency (TFIDF), document length, and spelling content-based features.

Content-based features are overwhelmingly used across our sample, with as many as 41 papers utilising con-tent-based features. As cyberbullying messages are often abusive and insulting in nature, it is not surprising that profanity was

found to be the most used content-based feature across the reviewed studies,with 22 papers using the presence of profanity in text as an indicator for cyberbullying. Studies such as Dinakar et al. (2011), Perez et al.(2012), Kontostathis et al.(2013), Nahar et al.(2013)and Bretschnei-der et al.(2014), created profanity lexicons using wordlists compiled by the researchers or sourced from external libraries such as noswearing.com3 and urban dictionary.com.By equating the presence of profanity to cyberbullying, the use of profanity lexicon salone fails to consider other key aspects of cyberbullying such as repeti-tiveness and the presence of a power differential. Rafiq et al.(2015)similarly cautioned against the use of profanity as the only feature for cyberbullying detection and argued that not all use of profanity and cyber-aggression constitutes bullying. Studies such as Nahar et al.(2013), Dadvar et al.(2014), Bretschneider et al.(2014) and Nahar et al.(2013)incorporate do ther features such as pro nouns in close proximity to profanity, since such personalised abusive content is potentially more indicative of cyberbullying than abusive terms on their own. For example, the phase "the f**king train was delayed again" is definitely not cyberbullying although it contained profanity but "you f**king idiot" could be. While this is an improvement, the pronoun + profanity feature still suffers the same short com-ings as using profane terms alone.

Dinakar et al.(2011),often cited for the performance gain achieved by their label-specific binary classifiers over multi-class classifiers, achieved this improved performance by using domain-specific content features learned from training classifiers on a set of messages clustered on sensitive topics such as race, culture, sexuality, and intelligence to then detect bullying messages within each cluster.

While Yin et al.(2009) did not find n-grams very effective in their experiments, its use as a detection feature is still relatively popular amongst studies, including Dinakar et al. (2011), Xu et al.(2012a; b), Sood and Churchill (2012a; b),and Munezero et al. (2014).As TFIDF provides a measure of a word's importance to a document within a collection of documents, it can sometimes provide better results than using n-grams in isolation (Yin et al., 2009). It is, therefore ,often used alongside n-gram and other features to im-prove detection performance ,as can be seen in the works of Yin et al.(2009), Dinakar et al. (2011), Dadvar and De Jong (2012), Sood and Churchill (2012a), and Nahar et al.(2013).

Of the 41 studies using content-based features, 5checked for the presence of cyberbullying keywords as part of the detection process. By cyberbullying keywords, we refer to non-profane words the use of which can indicate the presence of cyberbullying. These often are words associated with themes such as race, physical appearance, gender, and sexuality. As far back as the earliest study we discovered (i.e.,Mahmud et al., 2008), cyberbullying key-words have been used as detection features and this trend has continued with later studies such as Dinakar et al. (2011), Sanchez and Kumar (2011), Perez et al.(2012) and Dadvaret al.(2013b).These studies created lexicons com-posed of words so selected because their presence within a message or a post connotesa high likelihood of cyberbullying. For example, both Dinakaret al.(2011)identified themes such as race, culture, sexuality, physical appearance ,and intelligence as common bullying topics and used a lexicon of words associated with these themes as features ,while Sanchez and Kumar (2011) concentrated on homophobic slurs such as "gay", "queer", "homo" and "dyke" as keywords.[15]

3.1.2 Non textual Based:

While the focus of the studies in our sample has largely been on textual bullying, images and videos can also be used as delivery systems for online bullying and their im-pact can be as,or perhaps even more, damaging. In addition, as social media platforms improve their ability to detect and prevent textual bullying, bullies may likely resort to the use of other media forms to bypass anti-bullying measures. Recent advances in image processing and OCR (Optical Character Recognition)make it viable to attempt cyberbullying detection within media forms like images, animations ,and videos .With social media trends such as internet memes and viral videos becoming hugely popular in recent times, the scan be easily perverted by bullies to perpetrate cyberbullying. We, therefore ,envisage that developing systems capable of detecting bullying content within multimedia files is a key area for future research considerations. [15]

IV. Approaches of Models

4.1. Naive Bayes Model

4.2. SVM Model

4.3. DNN Model

4.1 Naive Bayes Model:

The Naive Bayes family of classifiers are simple conditional probabilistic classifiers that work by applying Bayes theorem with naive independence assumptions between the different features. All features are assumed independent given label Y:

$P(XI, \dots, Xn/Y) = \prod_{i=1}^{n} P(Xi/Y)$

A very simple document representation is used here, usually bag of words. Words important to the meaning of the text, and thus imperative in its classification, are considered, and given weight according to meaning, or in this case, severity. For instance, ''faggof" would receive a higher weight than "bitch", due to the former being sexually discriminatory and abusive.

Thus, given a document 'd' and class 'c':

$$P(c/d)=\frac{P(\frac{d}{c})+P(c)}{P(d)}$$

The maximum posterior class, or the most likely class, being in our case either bullying or not, would be:

C_{map} = $\text{argmax}_{c\text{EC}}$ P(c\d)

$$= \text{argmax}_{c\text{EC}} \frac{P(\frac{d}{c})+P(c)}{P(d)}$$

$=\text{argmax}_{c\text{EC}}$ P(d \c)P(c)

The corpus of data obtained to experiment with is the same as that used for J48. In this case, a true positive rate of 0.723, taking into account both textual and social features, was obtained. Without taking into account social features, the rate was 0.584 once again proving, as with similar tests performed with J48, that social features help improve the result.[18]

4.2 SVM Model:

SVM (Support Vector machine) is a supervised learning algorithm, and is one of the most efficient and universal classification algorithms. Its goal is to fmd the optimal separating hyperplane which maximizes the margin of training data. Initially the classifier is trained with labelled data before being used to classify the data to test accuracy. Before the data can be used to train our classifier, it is imperative to process it. This consists of the following steps:

• Labelling of data

• Generation of vocabulary

• Creation of document-term matrix

Once the labelled data is converted into a data matrix based on the values in the vocabulary, the values are then plotted and optimal hyperplane is chosen based on the convex hull. The optimal hyperplane is chosen in such a way that it maximizes the margin of the training data. Once the classifier is trained the input data is passed to this classifier to segregate it into positive and negative instances of bullying. This input data for testing purposes is also converted into data matrix and this data matrix is passed to the classifier. SVMs use sophisticated statistical learning theory to overcome the curse of dimensionality

Instead of specifying the feature vector, kernel functions can be used to provide similarity between data points. There are various kernels that can be used with SVM namely,

• RBF kernel (Radial basis function)

• Linear kernel

• Gaussian kernel

Linear kernel is a special case of the RBF kernel, and works best when the number of features is very large. The linear kernel on data sets acquired from Myspace, Kongregate and Slashdot datasets were used. The datasets are available from the workshop on Content Analysis for the Web 2.0 . The datasets contain manually-labeled data from , which is used as a ground truth dataset. Data from 3 different social networking sites are included in the dataset: Slashdot (496 files, 140,000 comments total (one for each article)), Kongregate (12 files, 150,000 comments total (one for each chatroom)) and MySpace (16346 files, 380,000 comments each (one for each thread)). Kongregate, an online gaming site, provides user messages from chat logs. Due to inherent frustration when playing online games, as well as a textual way to reach opponents, aggression is common in the posts. Slashdot is a discussion-based social networking site wherein users broadcast messages to others. MySpace is a popular social networking website. Datasets are in the form of XML files each containing and describing a discussion thread with multiple posts. Each post was extracted as as angular data element. Each data element is considered as one document and indexed through the inverted file index, assigning an appropriate weight to each individual term. Applying Lib SVM using a linear kernel, followed by tenfold cross-validation gives a false positive of 28 in 294 instances and false negative of 12 in 10184 instances . The model used in this case is the weighted TF IDF model. Over sampling of the training cases was used to improve the training. SVM with linear kernel using unigrams gives an accuracy of 79.6% while with bigrams it gives 8l.3% leading to the conclusion that bigrams should be used with the SVM linear kernel. This conclusion was obtained after testing the above on twitter corpus data (1762 tweets - 39% labeled as bullying traces). Taking all this data into consideration our conclusion is that linear SVM in combination with bigrams gives the best possible accuracy.[18]

4.3 DNN Model

Deep Layered Network Architecture

Deep neural networks compose computations performed by many layers. Denoting the output of hidden layers by $\mathbf{h}^{(l)}(\mathbf{x})$, the computation for a network with L hidden layers is:

f(x)=f[a(L+1) (h(L)(a(L) (...(h(2)(a(2)(h(1)(a(1)(x)))))))].

Each *preactivation function* $\mathbf{a}^{(l)}(\mathbf{x})$ is typically a linear operation with matrix $\mathbf{W}^{(l)}$ and bias $\mathbf{b}^{(l)}$, which can be combined into a parameter θ:

a(l)(x)=W(l)x+b(l), a(l)(xˆ)=θ(l)xˆ,l=1 a(l)(hˆ(l-1))=θ(l)hˆ(l-1),l>1

The "hat" notation x^ indicates that 1 has been appended to the vector **x**. Hidden-layer activation functions $\mathbf{h}^{(l)}(\mathbf{x})$ often have the same form at each level, but this is not a requirement.

In contrast to graphical models such as Bayesian networks where hidden variables are random variables, the hidden units here are intermediate deterministic computations, which is why they are not represented as circles. However, the output variables y_k are drawn as circles because they can be formulated probabilistically.

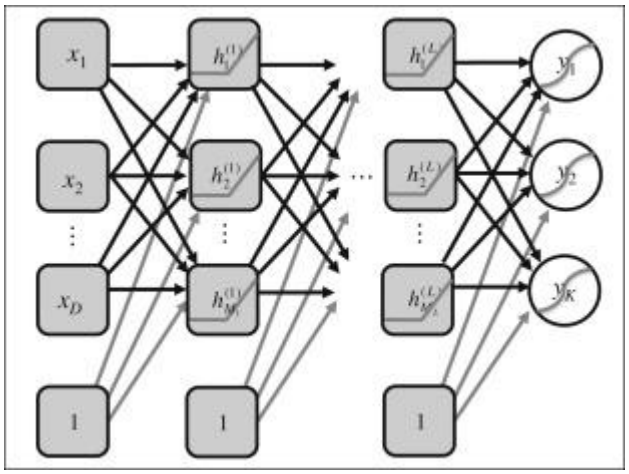

Figure 2: DNN Model[13]

V. Features

5.1 Detection of Non-Textual Cyberbullying

We are going to develop an application which has image in tweets or online data and we will fetch such image from twitter and after OCR classification will be done by our model SVM or naïve bayes model.

5.2 Expanding Cyberbullying Role Detection beyond Victims and Bullies

Roles such as instigators, defenders, and bystanders will be identified by us based on the algorithm model generated by us by collecting and labeling such type of data.

5.3 Determining a Victim's Emotional State after a Cyberbullying Incident

a victim may change his/her profile details following such interactions, post content containing negative sentiments, or leave the network abruptly. Such instigating interaction can be flagged up for subsequent review by a human who can then follow-up with appropriate actions Twitter will not allow to go in the profile of user for this we might create our own system which can identify such changes and will determine how the bullying affected person.

5.4 Word Representation Learning for Cyberbullying Detection

Experiments can be performed to generate word embeddings from different datasets, ranging from general corpora (e.g., Wikipedia) to more specialised datasets (e.g., abusive tweets) to compare their effectiveness for cyberbullying detection.

5.5 Detecting Cyberbullying in Streaming Data and Real-time

We will determine the cyberbullying on twitter dataset oauth token will be generated on twitter account we will fetch the tweets.

5.6 Evaluating Annotation Judgement

We will annotate the each twitter sentence and output will be generated shown on text. **[15]**

VI. Future Modification

The validity and accuracy of the predictive models to detect cyberbullying on twitter in this case primarily based on the correct psychometric categorization of the text.

In future it is intended to improve the system developed by use more accurate dataset and to detect the cyberbullying or not. We also apply other machine learning algorithm and check the accuracy of models. Higher accuracy model will help to detect more accurate bullying.

Another interesting direction for future work would be the detection of fine-grained cyberbullying categories such as threats, curses and expressions of racism and hate. When applied in a cascaded model, the system could find severe cases of cyberbullying with high precision. This would be particularly interesting for monitoring purposes. Additionally, our dataset allows for detection of participant roles typically involved in cyberbullying.

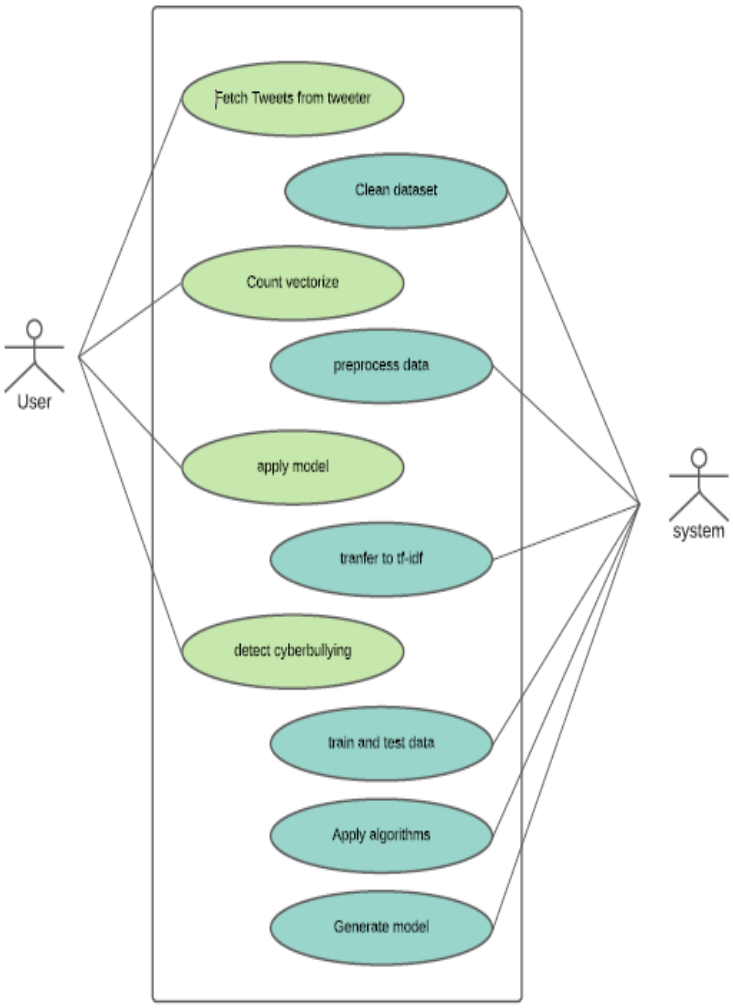

Figure 3: Interface of system

VII. Conclusion

The goal of this project is to the automatic detection of cyberbullying-related posts on social media. Given the information overload on the web, manual monitoring for cyberbullying has become unfeasible. Automatic detection of signals of cyberbullying would enhance moderation and allow to respond quickly when necessary. However, these posts could just as well indicate that cyberbullying is going on. The main aim of this project is that it presents a system to automatically detect **signals of cyberbullying** on social media, including different types of cyberbullying, covering posts from bullies, victims and bystanders.

VIII. References

[1] D. Poeter. (2011) Study: A Quarter of Parents Say Their Child Involved in Cyberbullying. pcmag.com. [Online].Available: http://www.pcmag.com/article2/0,2817,2388540,00.asp

[2] J. W. Patchin and S. Hinduja, "Bullies move Beyond the Schoolyard; a Preliminary Look at Cyberbullying," Youth Violence and Juvenile Justice, vol. 4, no. 2, pp. 148–169,2006

[3] Anti Defamation League. (2011) Glossary of Cyberbullying Terms.adl.org.[Online].Available:http://www.adl.org/education/curriculum connections/cyberbullying /glossary.pdf

[4] N. E. Willard, Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Research Press, 2007.

[5] D. Maher, "Cyberbullying: an Ethnographic Case Study of one Australian Upper Primary School Class," Youth Studies Australia, vol. 27, no. 4, pp. 50–57, 2008.

[6] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of Harassment on Web 2.0," in Proc. Content Analysis of Web 2.0 Workshop (CAW 2.0), Madrid, Spain, 2009.

[7] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," in Proc. IEEE International Fifth International AAAI Conference on Weblogs and Social Media (SWM'11), Barcelona, Spain, 2011.

[8] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. San Francisco, CA: Morgan Kauffman, 2005.

[9] R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kauffman, 1993.

[10] W. W. Cohen, "Fast Effective Rule Induction," in Proc. Twelfth International Conference on Machine Learning (ICML'95), Tahoe City, CA, 1995, pp. 115–123.

[11] D. W. Aha and D. Kibler, "Instance-based Learning Algorithms," Machine Learning, vol. 6, pp. 37–66, 1991.

[12] J. C. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," Advances in Kernel Methods, pp. 185–208, 1999. [Online]. Available: http://portal.acm.org/citation.cfm?id=299094.299105

[13]https://www.sciencedirect.com/topics/computer-science/deep-neural-network

[14]An Effective Approach for Cyberbullying Detection and avoidance ieee paper

[15]Approaches to Automated Detection of Cyberbullying: A Survey ieee paper

[16] Cyberbullying Detection System on Twitter ieee paper

[17] Methods for Detection of Cyberbullying: A Survey ieee paper

[18] Using Machine Learning to Detect Cyberbullying ieee paper

[19] Deep Learning Algorithm for Cyberbullying Detection ieee paper

[20] Online Social Network Bullying Detection Using Intelligence Techniques ieee paper