IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Cyber-Bullying Detection using Machine Learning Algorithms

Anvitha Keni[1], Deepa[2], Prof Mangala Kini[3],Deepika K V[4],Divya C H[5]

Abstract— Modern young people ("digital natives") have grown in an era dominated by new technologies where communications are pushed to quite a real-time level, and pose no limits in establishing relationships with other people or communities. The fast growing use of social networking sites among the teens have made them vulnerable to get exposed to bullying. Comments containing abusive words effect psychology of teens and demoralizes them. In this work we have devised methods to detect cyberbullying using supervised learning techniques. Cyber bullying is the use of technology as a medium to bully someone. Although it has been an issue for many years, the recognition of its impact on young people has recently increased. Through machine learning, we can detect language patterns used by bullies and their victims, and develop rules to automatically detect cyber bullying content. The data we used for our work was collected from the website kagle.com, it contains a high percentage of bullying content.

Keywords—cyber-aggressive; supervised; machine learning;

I. INTRODUCTION

SOCIAL Media 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 cyberbullying, which may have negative impacts on the life of people, especially children and teenagers. Cyberbullying can be defined as aggressive, intentional actions performed by an individual or a group of people via digital communication methods such as sending messages and posting comments against a victim. Different from traditional bullying that usually occurs at school during face to face communication, cyberbullying on social media can take place anywhere at any time. 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], cyberbullying victimization rate ranges from 10% to 40%. In the United States, approximately 43% of teenagers were ever bullied on social media [3]. One way to address the cyberbullying problem is to automatically detect and promptly report bullying messages so that proper measures can be taken to prevent possible tragedies.

To add up a social media called Twitter, Social media a powerful platform where you can have full freedom on what one wants to express or say; whether a negative or a positive one.

Suicide is the act of taking one's own life. Suicide is the second leading cause of death globally among people 15 to 29 years of age, according to the 2014 global report on preventing suicide by the World Health Organization [3]. Close to 800,000 people die due to suicide every year. For every suicide, there are more

people who attempt suicide every year. A prior suicide attempt is the most important risk factor for suicide in the general population. The age-standardized suicide rate in the Philippines is 5.8 for male, 1.9 for females, and 3.8 for both sexes. The rate is based from the number of cases affected per sample size of 100,000 people [2]. It is a misconception that suicide and depression affect mostly the poor. Stories abound of the growing prevalence of serious depression and suicide incidents in colleges attended by middle-class and rich kids [4].

Cyberbullying involves a person doing threatening act, harassment, etc. towards another person. Meaning of cyberbullying is a group(s) or an individual(s) of peoples that adopt telecommunication advantages to intimidate other persons on the communication networks [2]. However, most of the researchers in cyberbullying field take into account definition of cyberbullying. According to that, definition of cyberbullying formulated as "willful and repeated harm inflicted through the medium of electronic text".

Cyber bullying is when someone uses technology to send threatening Or embarrassing messages to another person. Bullying on social media can be even worse due to it's quick spread to the wider audience. Research shows that such behavior frequently occurred in Facebook and Twitter sites. It involves a person doing threatening act , harassment towards another person. Cyber bullying can takes into a few forms: lamming, harassment, denigration, impersonation, outing, boycott and cyber stalking.

A classifier is first trained on a cyberbullying corpus labeled by humans, and the learned classifier is then used to recognize a bullying message.The main roles involved in cyberbullying occurrences are cyberbully and victim. Given the aforementioned types of cyberbullying, there are various reasons why it happens. Apart from cyberbully and victim presences, proliferation of other roles may accentuate. According [10], they were classified the role of bullying into eight roles. These are of bully, victim, bystander, assistant, defender, reporter, accuser and reinforce.

A real-time sentiment analysis can be done using Big Data frameworks like Map Reduce and Hadoop. Hadoop framework shows a remarkable improvement in response time with a overall accuracy of 72% as compared to the classic Naive Bayes Classifier [13]. A combination of Feature vector including parameters like hashtags, emoticons etc.(ML) and knowledge-based approach was applied on Sanders analytics dataset. It consists of a total of 5600 tweets containing tweets of companies like Apple, Google and Microsoft [14]. Authors suggest that this advancement can be attributed to the use of hybrid approach.

RELATED WORK

In an effort to model the cyberbullying, Kelly Reynolds and April Kontosthatis, 2011[1] used machine learning to train the data collected from FromSpring.me, a social networking site, the data was labeled using Amazon Web service called Turk. The number of bad words were used as a feature to train model. In a study by Dinakar et al [2], states that individual topic-sensitive classifiers are more effective to detect cyberbullying. They experimented on a large corpus of comments collected from Youtube.com website. Ellen Spertus [3] tried to detect the insult present in comments, they used static dictionary approach and defined some patterns on socio-linguistic observation to build feature vector which had a disadvantage of high false positive rate and low coverage rate.

Altaf Mahmud et al [4] tried to differentiate between factual and insult statements by parsing comments using semantic rules, but they did not concentrate on comments directed towards participants and non- participants. Another work by Razavi et al [5] used a static dictionary and three level classification approach using bag- of-words features, which involved use of dictionary that is not easily available.

Dadvar et al., [7] analyzed the gender approach within the cyber bullying detection problem, applied to the social network MySpace, a platform that offers an interactive, user-submitted community of friends with personal profiles, blogs, groups, etc. Authors investigated the content of the posts written by the users but regardless of user's profile information. They used an SVM model to train a specific gender text classifier. The dataset consists of about 381.000 posts. The results obtained by the gender based approach improved the baseline by 39% in precision, 6% in recall, and 15% in F-measure.

At MIT, Dinakar et al. [4] applied different binary and multiclass classifiers on a manually labeled corpus of You Tube comments. This approach reached 66.7% of accuracy. Also, in this case authors used an SVM learner.

Xu, et al. [5] proposed different natural language processing techniques to identify bully traces and also defined the structure of a bully episode and possible related roles. Authors adopted Sentiment Analysis to identify roles and Latent Dirichlet Analysis to identify topics. Cyber bullying detection is formulated as a binary (positive/negative) classification problem and a linear SVM is trained with manually labelled dataset. The results reported 89% of cross validation accuracy, showing that even basic features and common classifier, can be useful to detect cyber bullying signals in text.

We can observe that most of these studies are based on supervised approaches, and usually adopt pre-trained classifiers to solve the problem, typically based on SVM. Data are manually labelled using online services or custom applications, and are usually limited only to a small percentage. NLP techniques are obviously wide adopted in all these works, due to the strict correlation between text analysis and cyber bullying detection. Mostly NLP tasks are performed at the preprocessing stage.

For English language, a notable amount of research have been performed in text categorization or cyber bullying detection. The research included YouTube comments, each was manually labelled and then various binary and multiclass classifications were implemented. Among the different classification techniques, SVM gets notable attention due to better performance in various text classifications. A recent study reported that the NB Classifier can be used effectively for Indian text classification. Researchers showed that classification results of SVM was better than the NB method for Urdu language. Hence the aim of this research was to explore various machine learning algorithms.

PROPOSED METHOD

Twitter dataset may easier to extracted compared to other mediums such as Facebook , Instagram, and YouTube. Even though statistic brain. come aforementioned stated that cyber bullying occurred most in Facebook but only data from public profiles could be extracted easily such as Twitter that the data is publicly available. The main function was to extract social media public data using available API. Then next step is to data cleaning and Pre-processing. As the extracted data had

multilingual unstructured content along with lot of emoji, it was required to clean the data for higher accuracy. Several supervised machine learning algorithms were compared to identify the best one.

Frequent use SVM by researches shows that SVM is popular among other classifiers in supervised learning approach. SVM is suitable for high-skew text classification such as to detect cyber bullying using content based features. Any circumstances such as missing data, type of feature and computer performance, SVM still perform other classifier.

These features are generally obtained by statistical analysis of documents (tweets or sentences):

Bad words:

From literature is quite evident and intuitive that some "bad" words make a text a suitable candidate to be labeled as a possible cyber bullying sentence. As just done in other works, and in this work we have identified a list of insults and swear words (550 terms), collecting these terms from different online available sources.

Bad words density:

In this work we check also the density of "bad" words as a single feature. This features is equivalent to the number of bad words that appear in a sentence, for each severity level, divided by the words in the same sentence.

Badness of a sentence:

We also add a feature to our work in order to measure the overall "badness" of a text. This feature is computed by taking a weighted average of the "bad" words (weighted by a severity assigned).

Density of upper case letters:

This feature is based on Dadvar et al. [7] results. The presence of capital letters in a text message is selected as a feature, considering it as possible 'shouting' at someone behavior, as commonly treated in social networks netiquette. This feature is given by the ratio between the number of upper case letter and the length (number of chars) of the whole sentence.

Exclamations and questions marks:

Just like capital letters, also exclamation points and question marks can be considered as emotional comments. We just stated that cyber bullying is related to an extreme case of sentiment analysis and so it can be connected to the strong (usually bad) emotions. With this premise, we consider helpful to introduce the number of exclamation points and question marks as a feature in our work.

The preprocessing step is done in the following:

- Tokenization: In this part we take the text as sentences or whole paragraph and then output the entered text as separated words in list.
- Lowering text: This takes the list of words that got the out of tokenization and then lower all the letters Like: "THIS IS AWESOME" is going to be 'this is awesome'.
- Stop words and encoding cleaning: This is an essential part of the preprocessing where we clean the text from those stop words and encoding characters like '*' which do not provide a meaningful information to the classifiers.

PROBLEM STATEMENT

We are interested in building a software application for detecting bullying instances in twitter. We are focusing on bulling detection for social reasons, which range from "reducing the number of suicides caused by bullying," to "making micro-blogging a no bullying zone"

METHODOLOGY

Automatic solutions related to cyber bullying detection are not properly studied in the past. This is one of the main reason for which there exists insufficient training datasets available. Some datasets are available instead on general sentiment analysis and all of them are used in supervised approaches. Although bullying messages are posted every day compared to hundreds of thousands of messages posted every second, they are very sparse. Collecting enough training data is an actual big challenge since random sampling will lead only to few bully messages. We selected two distinct datasets, recently published, related to the social network FormSpring.me and YouTube.

Dataset Collection & Preparation

For pursuing the task of sentiment analysis, dataset avail- ability is important. The performance of a classifier depends on the quality and size of dataset. Several datasets are available online [22] [23]. However to perform real time analysis, real time data is required. Such data can be collected using the Twitter API. Once the dataset has been prepared, it has to be split into training and test datasets.

Tweet Preprocessing and tokenization

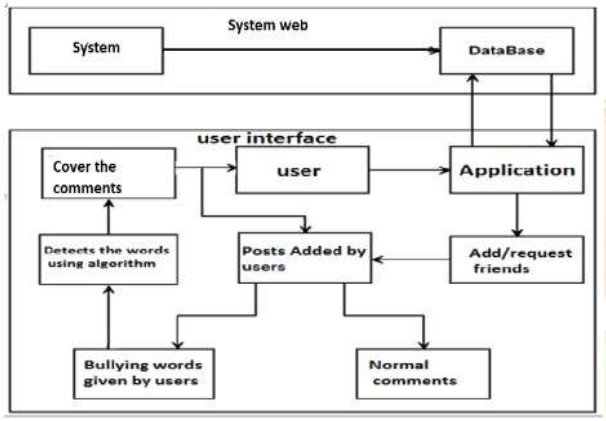
Twitter data is noisy. It contains emoticons, folksonomies, slangs and censored words [24]. Also it consists of short message texts. It is important to extract as much meaningful information as possible. Thus text pre- processing is a relevant step in performing sentiment analysis on Twitter data. The generic text pre-processing steps are listed below.

- Uppercase to Lowercase conversion. Replacement of URLs.
- Removal of extra symbols including punctuation marks and hashtags.
- Conversion of emoticons into the best suitable words.
- Stop words removal.
- Stemming of words into their root form.

Feature extraction and preparation of Feature vector

Machine learning-based automatic cyber-bullying detection involves two steps: Representation Learning for Tweets and Classification. Each tweet is converted into a fixed length vector. This constitutes the feature vector space. So higher the features, higher is the dimension of the feature vector space and this accounts to more processing and storage requirements.

This constitutes the feature vector space. So higher the features, higher is the dimension of the feature vector space and this accounts to more processing and storage requirements. This highlights the importance of Dimensionality Reduction in performing machine learning. Principle Component Analysis (PCA) and Latent Semantic Analysis (LSA) are methods to reduce dimensionality in feature vector space. After the preparation of vector space, the classifier can be trained on the training corpus. The classifier can be used to detect the presence of cyber-bullying in new tweets. Similar to other text classification tasks, the core step is numerical representation learning. Previous approaches utilize Bag-of-Words model to represent text. BoW model regards a document as a collection of words. Each document can be modelled as a vector whose weights correspond to the term frequency. Similar approach is the TF-IDF weighting scheme used in Information Retrieval Systems. BOW approach however, fails to capture the semantic information conveyed in sentences because it is indifferent to the order of words in a sentence.



Architecture Diagram

In This Architecture diagram we have two modules System Module and user Module.

In the system module we have system and database here the system is already pre-trained from kagle.com we will extracting the ready dataset. The system is directly connected to the database.

In User module using application user have to login and sign in using that application user can add friend and send friend request. If quotes added by the user is got any kind bulling words it will be sent to algorithm, here we are using the SVM algorithm there we are covering the comments with star symbol again it will sent to user and it is posted on the public if it is doesn't contain any bulling words it is considered as normal comment and it will sent to public .

Characteristics of Tweets

As pointed out in [1], we are aware of the many unique attributes of Twitter messages, which impose many challenges to our research with respect to the analysis of potential bullying instances. Some of these unique attributes are, to name a few:

1. Length. A maximum length of 140 characters per Twitter message. From our training set, the average length of a tweet was 16 words and the average length of a sentence was 88 characters.
2. Available data. We could easily collect huge amounts of data with ease. This is attributed to the fact that we are relying on public tweets and Twitter provides an API for collecting tweets.
3. Language Model. The use of misspellings, emoticons or other ASCII symbols, and slangs is highly frequent in tweets. So trying to decode them into something we could classify will be quite the challenge.
4. Domain. Tweet messages cover a variety of topics. That is, they are not tailored to a specific topic.

Data Gathering

Data was collected by using Twitter Advanced Search. In formulating the keywords, potential warning signs were searched and hints from psychological associations and organizations online. In addition, keywords from similar studies were borrowed and translated to Filipino language so that we can have a comprehensive data set which will cover majority of the Filipino twitter users, table 1.0 below shows the list of keywords used for data gathering. All in all, 5,174 tweets were collected in which 3,055 were English and 2,119 were Filipino or Taglish the data gathered was then saved in an excel file for data labeling.

Data Annotation

Data annotation was done by trained Psychologist and resident guidance counselor of the university. The classification was derived from the American Psychology Association's (APA) definition of suicide and its warning signs. The data is labeled with "0" and "1", 0 stands for the "non-risky" tweets and 1 for the "risky" tweets.

Classification Used in Cyberbullying Detection

Based on [12], authors implemented binary classifier; Naïve Bayes, Rule-based JRiP, Tree-based J48 and Support Vector Machine (SVM). Two experiments were set up where first experiment used binary classifier to train three labels dataset (intelligence, culture & race and sexuality). Second experiment was integrated three datasets into a single dataset and trained using multiclass. Result shows that binary classifier with three label were better in terms of accuracy instead of using multiclass classifiers with one dataset. Accuracy of rule-based JRip was better than other binary classifiers.

In other studied by [13] and [14], they mentioned binary classifier (SVM) used as classification algorithm. By integrated BoW and polarity, result for F-scores was better than using single feature in SVM classification.

On the other hand, [15] implemented linear SVM, logistic regression, decision tree and *AdaBoost* as classifiers. However, only linear SVM gave a better result instead of decision tree and *AdaBoost*. Result of precision and recall for both linear SVM and logistic regression were quite similar for cyberaggresion detection.

Linear SVM was also used by [16] to learn featured. EBoW model showed better performance in terms of precision and recall compared to BoW, sBow, LSA and LDA when implemented linear SVM as classifier

In summary, all of the studies used SVM as classification algorithm. Frequent use SVM by researchers

shows that SVM is popular among others classifiers in supervised learning approach. SVM is suitable for high-skew text classification such as to detect cyberbullying using content-based features [18]. Any circumstances such as missing data, type of features and computer performance, SVM still outperform other classifiers [17]. Table 1 shows the summarization of data source, features used and classification in cyberbullying detection for each research works as discussed.

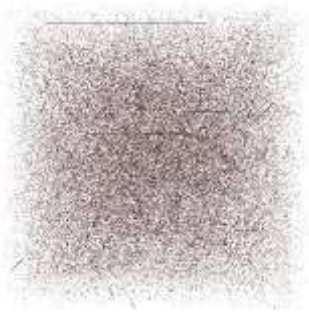
Study	Data Source	Feature	Classification
[13]	*Ask.fm*	• Content-based feature (BoW) • Sentiment-based feature (Polarity)	SVM
[14]	*Ask.fm*	• Content-based feature (BoW) • Sentiment-based feature (Polarity)	SVM
[16]	*Twitter* (http://research.cs.wisc.edu/bullying/data.html)	• Content-based feature (BoW, Bullying) • Latent semantic feature	Linear SVM

CHALLENGE IN CYBERBULLYING DETECTION

Results

As part of the evaluation of our results, we explored the use of Amazon's Mechanical Turk (Crowd sourcing) to classify unlabeled data, and to verify and validate newly labeled data. The use of Crowd sourcing along with that of machine learning algorithms helped us with the building of an infrastructure that could detect bullying instances in Twitter‟s public timeline. Table 4 summarizes the results of some of our experiments.

Despite the surfeit of data collected, the Twitter system would identify, target, and delete spammer (potential bullies) faster than we could fully inspect their social graphs. The resulting missing clusters of information would render a chaotic bullying graph; prohibiting any further analysis. Interestingly enough, one can interpret this observation as a sign of Twitter‟s success in implementing a powerful spam filter technology. Figure 3 shows one of resulting bullying graphs.



Bullying Graph

Evaluation parameters

Following parameters are used to compare the individual algorithms on the test datasets.

Recall:

The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn is the number of false negatives. It is the ability of the classifier to find all the positive samples. From Table IV we can observe that using features obtained using our hypothesis increases the recall value. This show that comments which are true bullied are predicted as bullied.

Precision:

The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp is the number of false positives. It is the ability of the classifier not to label a sample as positive that is negative.

AUC (Area under the curve score):

Computes the Area Under the Curve (AUC) from prediction scores and this evaluation parameter is strictly restricted to binary classification. As our task is to classify the comments as a bully/not bully i.e. a binary classification task, this evaluation parameter is very important. Using traditional feature extraction technique, model resulted in AUC score of 82%. An increase of 4% is achieved after introducing features extracted from our hypotheses

CONCLUSION

In this work tries to address the issue of cyber-bullying in Twitter platform using Machine Learning. Experiments were carried out with both supervised and unsupervised machine learning techniques. It was observed that identifying the right set of keywords is an essential step for getting better results during sentiment analysis Results indicate that our model achieves reasonable performance and could be usefully applied to build concrete monitoring applications to mitigate the heavy social problem of cyber bullying.

An experimental result indicates that the SVM based method achieves best accuracy and the performance improves if user specific data can be included. Due to high-dimensional input space, few irrelevant features and linearly separable nature of text dataset, SVM performs better than other classification algorithm for text classification. In future, significance of individual features can be studied for further enhancement of the method.

REFERENCE

[1] Rice, Eric, et al. "Cyber bullying perpetration and victimization among middle-school students." *American Journal of Public Health (ajph)*, pp. e66-e72, Washington, 2015.

[2] Bangladesh Telecommunication RegulatoryCommission, http://www.btrc.gov.bd/content/internet-subscribers-Bangladesh-january-2018, [Last Accessed on 18 Mar 2018].

[3] Mandal, Ashis Kumar,Rikta Sen. "Supervised learning methods for Bangla web document categorization." International Journal of Artificial Intelligence & Applications, IJAIA, Vol 5, pp. 5, 10.5121/ijaia.2014.5508

[4] Dani Harsh, Jundong Li, and Huan Liu, "Sentiment Informed Cyberbullying Detection in Social Media" *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Spinger, Cham, 2017

[5] Dinakar, Karthik, Roi Reichart, and Henry Lieberman. "Modeling the detection of Textual Cyberbullying*." The Social Mobile Web* 11.02(2011):11-17

[6] K. Dinkar, R. Reichart and H. Liebernman, "Modeling the Detection of Textual Cyberbullying," MIT. International Conference on Weblog nd Social Media. Barcelona, Spain, 2011.

[7] M. Dadvar and F.de Jong. 2012."Cyberbullying detection:astep toward a safer internet yard". In Proceedings of the 21st International Conference on World Wide Web(WWW '12 Companion). ACM, New York, NY, USA, 121-126

[8] Sunil B. Mane, Yashwanth Sawant, Saif Kazi, Vaibhav Shinde, "Real Time Sentiment Analysis of Twitter Data Using Hadoop", International Journal of computer Science and Information Technologies,(3098-3100),Vol.5(3),2014.

[9] Riya Suchdev, Pallavi Kotkar,Rahul Ravindran, "twitter Sentiment Analysis using Machine Learning and Knowledge-based Approach", International Journal of Computer Applications(0975-8887),Volume 103 a No.4, October 2014.

[10] J. Xu, K. Jun, X. Zhu, and A. Bellmore, "Learning from Bulling Traces in Social Media,*"Proc.2012 Conf.North Am. Chapter Assoc. Comput. Linguist. Hun. Lang. Technol, pp. 656-666,2012*

[11] S. Hnduja and J. W. Patchin *"Cyberbullying: Identification, Prevention, & Response,"Cyberbullying Res. Cent, no. October, pp.* 1-9,2018

[12] A. saravanaraj, J. I. sheebaassistant, S. Pradeep, and D. Dean, "Automatic Detection of Cyberbullying From Twitter." *IRACST-International J. Comput. Sci. Inf. Technol. Secur., vol. 6, no. 6,pp.* 2249-9555,2016.

[13] P. Badjatya, S. Guptha, M. Guptha, v. Varma, "Deep learning for hate speech detection in tweets*", Proceedings of the 26th International Conference on World Wide Web Companion, arXiv:1706.00188v1[cs.CL],June* 2017

[14] http://www.kdnuggets.com/datasets/index.html

[15] https://www.kaggle.com/datasets


IJCRT