IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CYBERBULLYING DETECTION USING MACHINE LEARNING

Shreenidhi B S[1], Mohammed Zaid Hulikatti[2], Nafey A H[3], Neha M R[4], Shradha S[5]

[2345] *Student [BE], Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management (DSATM), Bengaluru, Karnataka, India.*

[1]*Professor, Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management (DSATM), Bengaluru, Karnataka, India.*

Abstract: Cyberbullying has recently been reported to cause significant damage to society and the economy. Technological advances regarding the annotation of web documents and the diversity of online communities make the detection and tracking of such cases quite difficult and very difficult. Cyberbullying occurs through various information and communication technologies such as instant messaging, email, text messaging, social networking sites, blogs, and websites. We must fight this kind of abuse. The purpose of this project is to analyze and evaluate the communication between two specific individuals or an anonymous person. Recommended machine learning model to detect and prevent harassment in one-on-one conversation. Two classifiers, SVM and Naïve Bayes, are used to train and test social media bullying. Naive Bayes and SVM (Support Vector Machine) were able to detect true positives with 70% and 50% accuracy, respectively. SVM outperforms Naive Bayes by a similar dataset. We use the Twitter chat dataset to train the ml model.

***Keywords*— cyberbullying; classifiers; Naive Bayes; support vector machine (SVM); Machine Learning, Online Social Network, Social-Media, Text Classification**

I. Introduction

Cyberbullying means insulting, threatening, defaming, or intentionally harassing another person using modern means of communication, usually over an extended period of time. Cyberbullying occurs on the Internet (e.g. Via email, instant messaging, social networks, videos on various portals) or over the phone (e.g. Via WhatsApp or annoying calls). Most of the time, the perpetrator, known as the “bully,” acts anonymously, leaving the victim unaware of where the assaults take place. Especially in the case of cyberbullying between children and adolescents, the victim and abuser also know each other in the "real" world. Victims almost always suspect who may be behind the attacks. Cyberbullying often begins with people in the immediate environment: schools, neighbourhoods, villages or ethnic communities. Cases involving complete strangers are rare. Today, many young people are bullied. Bullies use various services like Twitter, Facebook, Email to intimidate people. Studies show that around 37% of children in India are involved in cyberbullying and almost 14% of bullying occurs regularly. Cyberbullying affects victims both emotionally and psychologically. Social media also allows bullies to exploit anonymity to satisfy their evil deeds. Things became even more serious when bullying incidents occurred continuously over time. So, preventing this from happening helps the victim. Cyberbullying is a constant invasion of privacy that doesn't stop at the doorstep (except when the new medium is not being used at home). The prevalence of information cannot be predicted because of the wide range of possibilities and the speed of new media. The perpetrators, so-called "cyber harassment," can act anonymously and often consider themselves safe because of this anonymity. In most cases, the identity of the perpetrator is presented very differently from the reality. Age or external image are not important criteria for cyberbullying. This can happen both between people of the same age (peers) and between people of different ages (student teachers). There is a possibility of unintentional cyberbullying, as thoughtless actions without awareness of the consequences can hurt the people involved. The abuser often fails to see these reactions and is unaware of the scale of the actions. Given the consequences of cyberbullying on its victims, it is imperative to find the appropriate actions to detect and prevent it. Machine learning is one of the successful approaches that learns from data and creates a model that automatically classifies appropriate actions. Machine learning can be useful to detect language patterns of bullies and thus can generate a pattern to detect acts of cyberbullying. A recent study conducted by Microsoft Corporation to understand the spread of cyberbullying globally showed that India ranks 3rd in terms of cyberbullying after China and Singapore. According to recent studies, 52% of young people in India have been victims of cyberbullying in the past and around 38% of them have been bullied. Cyberbullying is basically of two types, one that contains abusive language and one that embarrasses the intended target but does not use swear words. Posts containing abusive content or profanity are more likely to be labelled as “articles”. According to, for today's younger generation, "Gay", "Bitch" and "Slag" are the most commonly used abusive terms in schools.

Examples:
"Kevin is a faggot." (Openly abusive)
"Rohan looks good in a mini skirt." (No abusive language involved)

India has many cases of bullying. 79% of Indians are aware of and concerned about cyberbullying, compared to 54% globally. 53% of Indians have been bullied compared to the world average of 37%. On top of that, 50% of Indians have ever engaged in cyberbullying while globally only 24% of the population has been involved in similar incidents. In contrast, 63% of Indians are educated and 76% of organizations have a formal policy on cyberbullying, compared with global averages of 23% and 37%. The purpose of this article is to analyze and evaluate communication between two specific individuals or an anonymous person. A proposed machine learning model to detect and prevent harassment on chat interfaces. Two classifiers, SVM and Naïve Bayes, are used to train and test social media bullying. Naive Bayes and SVM (Support Vector Machine) were able to detect true positives with 70% and 50% accuracy, respectively.

II. Proposed System

In the recommendation system, abusive content is checked and reported by the machine, not other users. Some disrespectful words/slang was flagged as abusive and users suggested this as well. The bullied user will then be prompted to indicate whether they want to block and report the abuser. If the user chooses to block others, this is done immediately, otherwise the case is considered a false positive. A dataset used to train a computer to recognize abuse and distinguish it from other texts. The trained model is performed on a chat interface between two users. If there is an abuse/swearing event, the abused user will be asked whether to block them and take the necessary action. The entire approach to detecting and preventing harassment in the context of one-on-one dialogue against cyber behavior is divided into two main phases: model development and experimental setup.

1. Experimental setup:

Step-by-step process SVM and Naive Bayes are used to detect cyberbullying.

Steps:
1. For a specific location, a limited number of tweets are retrieved through the Twitter dataset.

2. Data pre-processing and data extraction will be performed on the retrieved Tweets.

3. Pre-processed tweets are forwarded to the SVM and Naïve Bayes model (see Model Development) which calculates the probability of retrieved tweets to verify if a retrieved tweet is worthwhile. afraid or not.

4. If the probability of a tweet being retrieved is between 0 and 0.5, the tweet will not be considered harassed. If the probability of the tweet being fetched is greater than 0.5, it will be added to the database and then another 10 tweets from that user's timeline will be fetched, since they cannot directly say that whether the person is bullying someone because it is possible that he is chatting with his friend to make sure he is bullying someone, we report the user. This particular may or may not have a history of similar flags.

5. Again, the list of tweets from the user's timeline is fed into the SVM and Naive Bayes model to predict the outcome of the tweets.

6. And again, the average probability of this user's tweets will be calculated and if it is greater than 0.5 it will be considered a bullied tweet and will be saved in their database I. If the mean probability is less than 0.5, the record will be deleted from the database.

The first step in the solution is to collect the tweets dataset. In the next two steps are data preprocessing and feature extraction is performed over the tweets. And after performing preprocessing and feature extraction tweets are passed to the SVM model for classification to predict whether the tweet is Bullying or Non-Bullying.

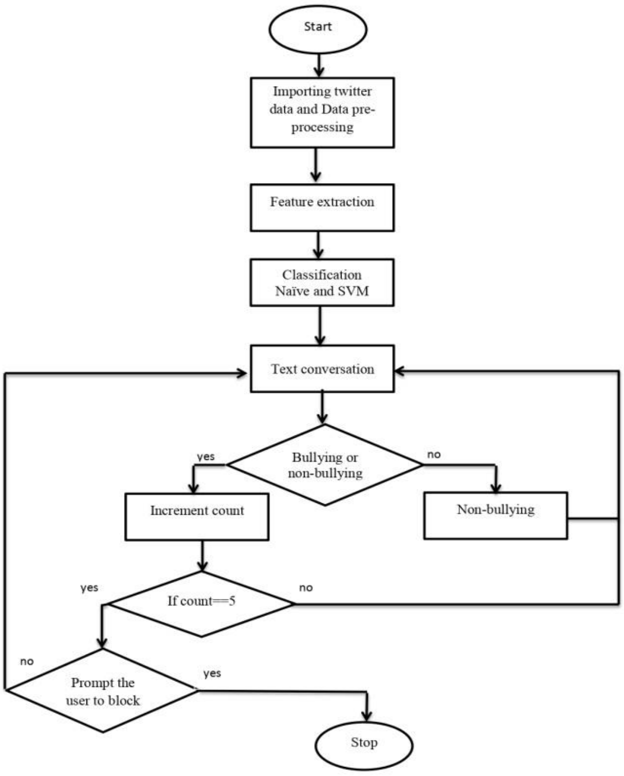


Fig. I show a diagram of the proposed solution.

2. Developing the Model:

The entire model is divided into 3 major steps: Pre-processing, the algorithm, and feature extraction.

A. Pre-processing:

The Natural Language Toolkit (NLTK) is used for the Pre-processing of data. NLTK is used for tokenization of text patterns, to remove stop words from the text, etc.

- Tokenization: In tokenization, the input text is split as the separated words and words are appended to the list. Firstly, PunktSentenceTokenizer is used to tokenized text into the sentences. Then 4 different tokenizers are used to tokenize the sentences into the words:
  - o WhitespaceTokenizer
  - o WordPunctTokenizer
  - o TreebankWordTokenizer
  - o PunctWordTokenizer
- Lowering Text : It lowers all the letters of the words from the tokenization list. Example: Before lowering "Hey There" after lowering "hey there".
- Removing Stop words: This is the most important part of the preprocessing. Stop words are useless words in the data. Stop words can be get rid of very easily using NLTK. In this stage stop words like \t, https, \u, are removed from the text.
- Wordnet lemmatizer: Wordnet lemmat izer finds the synonyms of a word, meaning and many more and links them to the one word.

B. Feature Extraction:

In this step, the proposed model has transformed the data in a suitable form which is passed to the machine learning algorithms. The TFDIF vectorizer is used to extract the features of the given data. Features of the data are extracted and put them in a list of features. Also, the polarity (i.e. the text is Bullying or Non-Bullying) of each text is extracted and stored in the list of features.

C. Algorithm Selection:

To detect social media bullying automatically, supervised Binary classification machine learning algorithms like SVM with linear kernel and Naive Bayes is used. The reason behind this is both SVM and Naive Bayes calculate the probabilities for each class (i.e. probabilities of Bullying and Non-Bullying tweets). Both SVM and NB algorithms are used for the classification of the two-cluster. Both the machine learning models were evaluated on the same dataset. But SVM outperformed Naive Bayes of similar work on the same dataset. Classification reportis also evaluated. The accuracy, recall, f-score, and precision are also calculated.

Precision = TP / (TP+FP)
Recall =TP/(TP+FN)
F-Score = 2*(Precision*Recall) / (Precision + Recall)
Where TP = True positive numbers
TN = True negative numbers
FN = False negative numbers
FP = False positive numbers

Iii. Literature survey

Detecting A Twitter Cyberbullying Using Machine Learning [1]. To identify word similarities in the tweets made by bullies and make use of machine learning to develop an ML model to automatically detect social media bullying action. The goal of this paper is to show the implementation of software that will detect bullied tweets, posts, etc. A machine learning model is proposed to detect and prevent bullying on Twitter. Two classifiers i.e. SVM and Naïve Bayes are used for training and testing the social media bullying content.supervised Binary classification machine learning algorithms like SVM with linear kernel and Naive Bayes is used to detect the true positives with 71.25% and 52.70% accuracy respectively. Also, Twitter API is used to fetch tweets and tweets are passed to the model to detect whether the tweets are bullying or not. The reason behind this is both SVM and Naive Bayes calculate the probabilities for each class (i.e. Probabilities of Bullying and Non-Bullying tweets).

The proposed model is divided into 2 major steps: 1) Experimental Setup 2) Developing the model

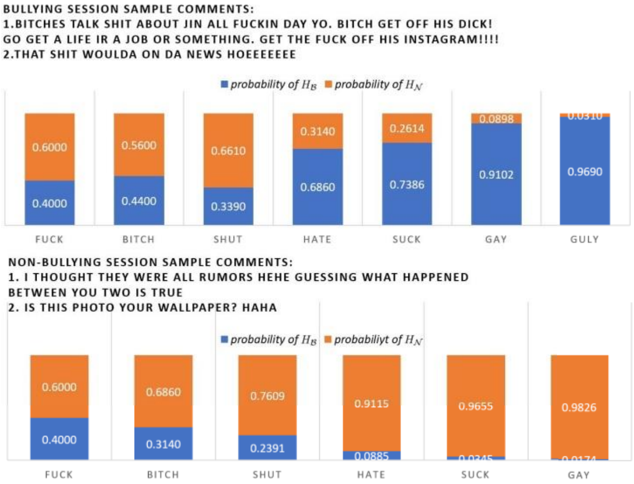
* Experimental setup
  * Twitter API extraction
  * Data pre-processing
  * Calculating probability.

* Developing the model
  * Preprocessing
    * Tokenisation
    * Lowering text
    * Removing stop words
    * Word net lemmatizer.
  * Feature Extraction
  * Algorithm selection.

Automation in Social Networking Comments With the Help of Robust fastText and CNN.[2] .The proposed model in paper aims to apply the text-based Convolution Neural Network (CNN) with word embedding, using fastText word embedding technique. fastText has shown efficient and more accurate results compared to Word2Vec and GLOVE model. The proposed model aims to improve detecting different types of toxicity to improve the social media experience. This model classifies such comments in six classes which are Toxic, Severe Toxic, Obscene, Threat, Insult and Identity-hate. Multi-Label Classification helps us to provide an automated solution for dealing with the toxic comments problem people are facing.

Cyberbullying Detection on Instagram with Optimal Online Feature Selection [3]. This focus on Instagram as the online social media platform with the highest percentage of users reporting experiencing cyberbullying. Instagram is a social media platform that allows users to share pictures and videos either publicly, or privately to their followers. This paper formulate cyberbullying detection as a sequential hypothesis testing problem, and propose a novel algorithm designed to reduce the time to raise a cyberbullying alert . The paper minimize the number of feature evaluations necessary for a decision to be made .The algorithm sequentially reviews features starting from the most informative, and decides when to stop. A key property of this solution is that in accomplishing these goals, it does not adversely impact classification quality

Abdullah-Al-Mamunet al.[4] has trained a machine learning model to detect social media bullying for Bangla text.
The data extraction from the datasets obtained from twitter API ,Facebook graph API specific to Bangladeshi text .
The trained various machine learning model for different algorithm for cyberbullying and sentimental analysis on Bangla text. The further enhanced there study to detect sentimental analysis not only on the Bangladeshi text but to English conversation. They have used supervised Machine Learning algorithms i.e. Support vector machine, k-nearest neighbor, and NB (Naive Bytes) classifier models. Accuracy of every Model:
SVM-97.27% | KNN-96.73 |
NB- 97.23SVM outperformed other classifiers for both English and Bangla text.



BULLYING SESSION SAMPLE COMMENTS:
1.BITCHES TALK SHIT ABOUT JIN ALL FUCKIN DAY YO. BITCH GET OFF HIS DICK!
GO GET A LIFE IR A JOB OR SOMETHING. GET THE FUCK OFF HIS INSTAGRAM!!!!
2.THAT SHIT WOULDA ON DA NEWS HOEEEEEEE

NON-BULLYING SESSION SAMPLE COMMENTS:
1. I THOUGHT THEY WERE ALL RUMORS HEHE GUESSING WHAT HAPPENED
BETWEEN YOU TWO IS TRUE
2. IS THIS PHOTO YOUR WALLPAPER? HAHA

fig. 1. posterior probability evolution as more features are examined for a cyberbullying (upper plot) and a non–cyberbullying session (lower plot)

IV. Result

The two algorithms support vector machine and naive Bayes are applied on the collected dataset from Kaggle, github, etc. are compared. After preprocessing the data and feature extractions on the dates for training and testing are divided. The SVM and Naive Bayes are evaluated to calculate the accuracy. SVM outperformed Naive Bayes.

The Support vector machine got the accuracy of 72 percent and the naive Bayes approach show 53 percent accuracy. The results of the application developed by the paper is fetching the twitter dataset and trying the model and to classify it into bullying or not depends on the one-to-one conversation and fetching the conversation and classifying it into bullying or not with their probability for every 5th tweet flagged as "Bullying" conversation and rest are "Non-bullying". The bullied person is presented with the option to block the user automatically if the abbused person decides if he is being bullied then the abuser os blocked from the conversation.

Classifiers	Accuracy (in %)
Naive Bayes	52.70
Support Vector Machine	71.25

V. Future enhancement

This model can be implemented using behavior analysis. Ergo, the abuser's texting patterns are learnt by the machine. Based on the learning the abuser will be blocked right away from the platform.

If the abuser has a history of similar abusive behavior, then on a count of 10 such abusing occurrences, the user will be banned from the platform.

If the user is blocked from the platform and tries to create a new account/profile on the same platform using the same credentials (one or more), then the user will not be allowed to create the account at the very beginning.

The mac-address of the abuser is traced and that particular mac-address is blocked from the domain of the platform.

VI. Conclusion

This study reviewed existing literature to detect aggressive behavior on social media. By using various machine learning approaches. We have reviewed four main features of Detecting cyberbullying messages by using machine learning approaches namely Importing data, feature extraction, construction of cyber bulling detection model and evaluation of the constructed model. An approach is proposed for detecting and preventing Cyberbullying using supervised Binary classification machine learning algorithms.

Model is evaluated on both Support Vector Machine and Naive Bayes. The datasets used in this is a collection of a tweets that have been classified as positive, negative or neutral bullying. Before training and testing, collected set of tweets go through several phases of cleaning, Normalization, tokenization and feature selection.

In the data analysis, we keep a count of number of abusive words a particular person as an indication of aggressive behavior. Results shows that the accuracy for detecting cyberbullying content with Naive Bayes has a great for support of around 70% which is better than support vector machine. So as a result of that the most effective supervised machine learning classifiers for classifying cyberbullying messages were identified. We have used accuracy, precision recall and f-measure which gives us The curve function for modelling the behavior in cyberbullying.

Our model will try and help in reducing cyberbullying to some extent.

VII. Acknowledgement

It gives us immense pleasure to present before you our paper titled "CYBERBULLYING DETECTION USING MACHINE LEARNING". The joy and satisfaction that accompany the successful completion of any task would be incomplete without the mention of those who made it possible. We are glad to express our gratitude towards our prestigious institution DAYANANDA SAGAR ACADEMY OF TECHNOLOGY AND MANAGEMENT for providing us with utmost knowledge, encouragement and the maximum facilities in undertaking this project.

We would like to express my deepest gratitude and sincere thanks to our guide Prof. Shreenidhi B.S, Department of Computer Science and Engineering, DSATM, Bangalore, for his keen interest and encouragement in the project idea whose guidance made this technical paper into reality.

We are grateful to our coordinator Dr. Kavitha C, Professor of Department of Computer Science and Engineering, DSATM, Bangalore, for his valuable suggestions and advice throughout our work period.

We express our deepest gratitude and special thanks to Dr. C. Nandini, Vice Principal & Head of Department Computer Science Engineering, for all her guidance and encouragement.

We wish to express a sincere thanks to our respected Principal Dr. M Ravishankar for all their support.

VIII. References

[1] Detecting A Twitter Cyberbullying Using Machine Learning Rahul Ramesh Dalvi, Sudhanshu Baliram Chavan, Aparna Halbe Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020).IEEE Xplore Part Number: CFP20K74-ART; ISBN: 978-1-7281-4876-2

[2] Automation in Social Networking Comments With the Help of Robust fastText and CNN" .Mestry, Suresh; Singh, Hargun; Chauhan, Roshan; Bisht, Vishal; Tiwari, Kaushik (2019). [IEEE 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT) - CHENNAI, India (2019.4.25-2019.4.26)]

[3] Cyberbullying detection on instagram with optimal online feature selection.Mengfan yao, charalampos chelmis, daphney–stavroula zois 2018 ieee/acm international conference on advances in social networks analysis and mining (asonam)

[4] Abdhullah-Al-Mamun, Shahin Akhter, "Social media bullying detection using machine learning on Bangla text", 10th International Conference on ElectricalandComputerEngineering,pages385-388,IEEEXplore,20

[5] Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges (Mohammed Ali Al-garadi,Mohammad Rashid Hussain, Henry Friday Nweke)

[6] https://stanford-cs221.github.io/spring2021/

[7] https://reactjs.org/tutorial/tutorial.html

[8] https://www.tensorflow.org/tutorials

[9] http://introtodeeplearning.com/

[10] Natural Language Processing in Action Understanding, analyzing, and generating text with Python - Hobson Lane Cole Howard Hannes Max Hapke

[11] Introduction to Machine Learning with Python -Andreas C. Müller & Sarah Guido

[12] Machine Learning – Tom M. Mitchel

[13]https://muthu.co/understanding-the-classification-report-in-sklearn/

[14] https://developer.twitter.com/en/apps

[15] https://text-processing.com/demo/tokenize/

[16] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[17]https://towardsdatascience.com/naive-bayes-classifier-81d512f50ac

[18] Phillips, L., et al., Using Social Media to Predict the Future: A Systematic Literature Review. arXiv preprint arXiv:1706.06134, 2017.

[19] Quan, H., J. Wu, and J. Shi, Online social networks & social network services: A technical survey. Pervasive Communication Handbook. CRC, 2011: p. 4.

[20] Peterson, J.K. and J. Densley, Is Social Media a Gang? Toward a Selection, Facilitation, or Enhancement Explanation of Cyber Violence. Aggression and Violent Behavior, 2016.

[21] BBC, Huge rise in social media 'crimes' http://www.bbc.com/news/uk-20851797, 2012

[22] Abdhullah-Al-Mamun, Shahin Akhter, "Social media bullying detection using machine learning on Bangla text", 10th International Conference on Electricaland Computer Engineering, pages385-388, IEEEXplore, 2018

[23] Cyberbullying Detection Using Social and Textual Analysis - Qianjia Huang, Vivek Singh, Pradeep Atrey

[24]http://download.rnicrosoft.com/download/E/8/4/E84BEEAB-7B92-4CF8-B5C77CC20D92B4F9/WW%200nline%20Bullying%20Survey%20- %20Executive%20Summary%20-%20SingaporeJinal.pdf (Accessed 31st August)

[25] www.youtube.com

[26] www.facebook.com

[27] www.myspace.com

[28]www.instagram.com