Cyber-Bullying Detection Using Machine Learning and Naïve Bayes and N-Gram Model

Prof. Gauri Rao[1], Mehul Goyal[2], Diksha Wali[3], Sarthak Yadav[4]
[1,2,3,4]*Department of Computer Engineering, Bharati Vidyapeeth Deemed to be University College of Engineering Pune, India*

Abstract **- Many social media channels such as Facebook, Twitter, Instagram etc has altered our lives. People are mow connected to the world via these social media channels. These social media platforms have remarkable features but have their disadvantages too. Communicating through social media from a remote location and without revealing the real identity has given birth to a new crime that is CYBERBULLYING. Cyberbullying is basically misuse of this technology to tease, insult, harass or humiliate a person through internet. Many attempts have been introduced to prevent and decrease the number of cyberbullying cases however these methods rely on the interaction with the victim hence there is a need of method for cyberbullying detection where there is no involvement with the victim. In this paper we have reviewed and analysed existing models and propose a method for cyberbullying detection using Naïve Bayes Classification and N-Gram Model to scrutinize the bullying scenario or sentiment from each and every tweet collectively.**

Index Terms **- Naïve Bayes Classification, N-Gram Model, Text Mining, Sentiment Analysis.**

I.INTRODUCTION

In today's world every person is connected to each other via social media and various such platforms are Instagram, Facebook, Twitter etc. Today almost 1 out of 3 people use social media and its expected that there will be more than 3 billion active social media users globally by the end of 2021 Besides among these various social media platforms Twitter is an important platform as it is used as a vital data source for researches and Twitter is also used as a popular microblogging network operating in real time and news often appears in it before it appears in social sources.

With the increase in active users on social media the cases for cyberbullying has also increased rapidly. Cyberbullying is a form of violence that mainly occurs to children or adolescents and is done generally by their friends of their age through internet.

Cyberbullying is considered as crime because it is an incident when a child or teenager is teased, insulted, humiliated, or harassed by another child or teenager through digital technology.

Cyberbullying takes various forms and methods, and it includes threatening messages via e-mail, uploading inappropriate images of victims, creating websites to spread slander and poking fun at victims to accessing other peoples social networks accounts and threatening them and causing trouble. Using services like email and instant messengers it becomes easier for the bullies to do their nasty deeds without disclosing their identity.

We can differentiate between cyberbullying and the traditional bullying on the basis of the effect left on the victim. Traditional bullying leaves physical damage and also leaves emotional and psychological damages but in cyberbullying its all emotional and psychological.

II.MOTIVATION

The recent increase in use of social media and its rising active users has lead to a rise in cyber-bullying cases around the globe so we need an approach where with the help of machine learning we can detect language patterns and try stopping cyberbullying.

Many initiatives have been taken around the world to scrutinize cyberbullying like for example University of Turku, Finland has an anti-cyberbullying program called Kiva, an anti-harassment campaign in France, an anti-cyberbullying initiative by the Belgium Government.

However, despite these initiatives still there is a rise in cyberbullying cases and the reason being the internet's

content is very vast and it is quite difficult to control and analyse that data to filter out and detect cyberbullies.

Advancement in techniques from domains like machine learning, statistics and text recognition have made it possible to design an approach for intervention and stop cyber-bullying.

III LITERATURE SURVEY

As mentioned before there are many approaches suggested before in order to stop and control the cyberbullying cases. In this section we present a detailed description of such models presented in various research done before.

In [2017] Hatoon AlSagri and Mourad Ykhlef proposed a binary classification method which it identifies whether a person is depressed or not and it is done on the basis of his/her twitter tweets and twitter profile activity. Different Machine learning algorithms are used and various datasets are being used

In [2021] Reem Bayari and Ameur Bensefia conducted an in-depth analysis where 16 studies on automatic cyberbullying detection methods based on text language were explored. They also undertook various parameters in their study like features, language, dataset size and dataset source of the latest research in this field

In [2020] Amgad Muneer and Suliman Mohammed Fati proposed a cyberbullying detection model in which several classifiers based on TF-IDF and Word2Vec feature extraction are used and various text classification methods based on Machine learning were used and global twitter dataset was used for experimentation.

In [2017] Noviantho Sani Muhamad lsa and Livia Ashianti proposed optimal SVM kernel in classifying cyberbullying with an average accuracy of 97.11% as the data used here are non-linear separable hence the optimal function for separating the sample into different classes is SVM with Poly kernel.

In [2019] Amisha Akhter, Uzzal K. Acharjee and Md. Masbaul Alam Polash conducted a work for detection and classification of cyberbullying from comments in Facebook posts and bullies are classified into three categories – Shaming, Harassment and Racism. A Multinomial Naïve Bayes classifier classifies bully comments here.

In [2020] Shailvi Sharma and Dharmveer Singh conducted a research which showed a pattern of sentences that have the greatest potential in cyberbullying that is “Subject + Word Bully” and within the context it has the greatest potential used to carry the cyberbullying activities and there is a very marginal difference while using Naïve Bayes classification with various Gram Models comprising of Uni, Bi and N models.

In [2019] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer and Ammar Mohammed proposed an approach to detect cyberbullying using machine learning techniques and used two classifiers SVM and Neural Network and used TFIDF and sentiment analysis algorithms and the classifications were evaluated using different N-Gram Models and achieved 92.8% accuracy using Neural Network with Tri-Gram and 90.3% accuracy using SVM with 4-Grams while using TFIDF and sentiment analysis together.

IV APPROACH

In the proposed model we have 3 major stages:

1. In this technique first we collect the raw data. The log data tweet from the twitter is taken and stored in comma separated value
2. In the second stage we perform pre-processing and cleaning of the data to make data structured and easy to analyse
3. The last stage is the classification of the pre-processed data using Naïve Bayes Classification integrated with Uni-Gram, Bi-Gram Tri-Gram and N-Gram

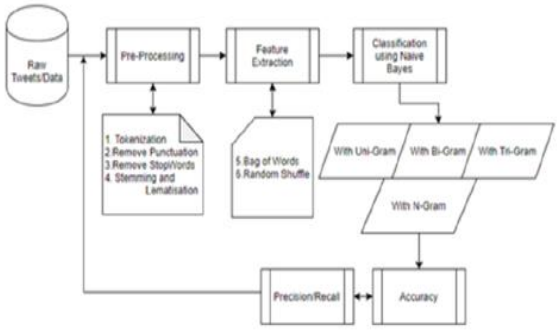

Fig.1 Stages of the Proposed Method

N GRAM

The N-Gram was initially designed by Russian Mathematicians in the early 20th century where items

can be either script/characters or words or others according to the appliance and then one of the word based n-gram models is used in predicting the next word in a certain word order. In other words the n-gram is just a container of words with each having length n.

1. An n-gram with size 1 is called Uni-Gram
2. An n-gram with size 2 is called Bi-Gram
3. An n-gram with size 3 is called Tri-Gram

And so on.

For Character generation the N-gram consists of substrings along the n characters of a string or in other words the N-Gram is a slice of the number of n characters from the string. It is used to take n-character pieces of a number of words that are continuously read from the source text to the end of the document.

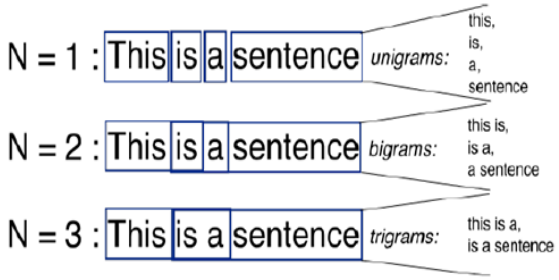

Fig.2 Various N-Gram Models

NAÏVE BAYES CLASSIFIER

It is an algorithm which is used to find the highest probability value to classify the data in a appropriate category.

Here our data or test data is the tweets documents and there are two stages of classification in this document

1. Stage 1 is the training on the documents with the known categories
2. Stage 2 is the classification of the document whose category is unknown

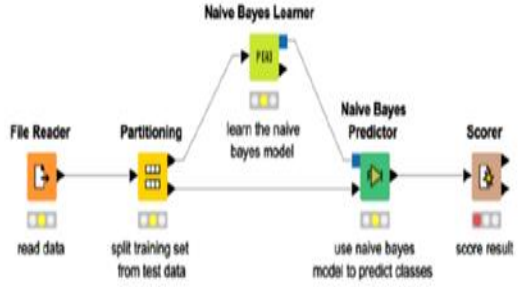

Fig.3 Naïve Bayes Classification Process

ACCURACY

At the end we try calculating the accuracy of our model to check with which N-Gram model our Proposed approach is more efficient

Accuracy is basically the ratio of number of correct predictions to the total number of detections

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Fig.4 Accuracy

Here

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

V. PROJECT PLANNING

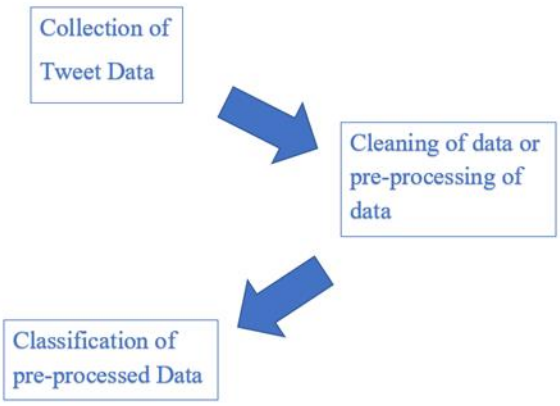

Fig.4 Stages of Proposed Method

As discussed our approach has 3 main steps which are mentioned in above figure. Here we will study about the whole process in detail.

In the first step we try collecting our data which will be our test data and we collect this data from Twitter

The second stage is data pre-processing and here its done to improve the quality of the research data and here various steps we perform are

1. Remove the stop words
2. Remove the Extra words
3. Remove the Hyperlinks

We also perform Feature Extraction in Pre-Processing that includes Noun, Adjective and Pronoun and the frequency of the words in the text is determined and recorded.

Then next the data is sent to the N-Gram Model which is used to predict the next word in a certain order taking n length words,

N-Gram takes n-character pieces of a number of words which are read continuously from the source data set text

For Example the word "Modem" is read different in different N-Gram Models:

1. In Uni-Gram – M, O, D, E, M
2. In Bi-Gram – MO, OD, DE, EM
3. In Tri-Gram – MOD, ODE, DEM

For the sentence the N-gram takes n-word pieces from the series of the words (sentence, paragraph, reading) which are continuously read from the source

For Example the sentence is "there is a car":

1. In Uni-Gram- there, is, a, car
2. In Bi-Gram- there is, is a, a car
3. In Tri-Gram- there is a, is a car

The benefit of using N-Gram is that we are not using whole word here which is sensitive to errors written in a document

After this we use Naïve Bayes Classification Model which helps in classifying the comments into two categories

1. Bullying
2. Non-Bullying

VI. CONCUSION

In the research its been successfully proved that there are words which have great potential and are used maximum time to carry out cyberbullying activities and it is known that the word "IDIOT" has the greatest potential

There was a common pattern in sentences which have greatest potential to carry out cyberbullying activities and that pattern is "Subject + Word Bully".

Classification of cyberbullying done using Naïve Bayes Classification and N-gram Model is able to achieve accuracy

1. Naïve Bayes + Uni-Gram is 67.12%
2. Naïve Bayes + Bi-Gram is 67.78%
3. Naïve Bayes + Tri-Gram is 55.98%
4. Naïve Bayes + Ni-Gram 66.45%

As a result we have seen very marginal difference while using Naïve Bayes with different N-Gram Models but integrating Naïve Bayes with Bi-Gram we got higher accuracy

REFRENCES

[1] Kumar, Uday, "Present scenario of cybercrime in INDIA and its preventions" International Journal of Scientific and Engineering Research. volume 6. 1971, 2015

[2] Tseng, Chris & Pateli, N. & Paranjape, Hrishikesh & Lin, T.Y. & Teoh, SooTee, "Classifying twitter data with Naïve Bayes Classifier", 2012

[3] Kaur, Simrat & Kalsi, Shaveta, "Analysis of Wheat Production using Naïve Bayes Classifier", International Journal of Computer Applications, 2019

[4] Oktaviana, Shinta & Ermis, Iklima & Anasanti, Mila & Hammad, Jehad, "Network Disruption Prediction Using Naïve Bayes Classifier", 2019

[5] Anyanwu, C & Udanor, Collins, "An N-Gram Determination of Twitter User Sentiments", 2020

[6] Kansara, Krishna B. and Narendra M. Shekokar, "A Framework for Cyberbullying Detection in Social Network.", 2015

[7] Setty, Shankar & Jadi, Rajendra & Shaikh, Sabya & Mattikalli, Chandan & Uma, M, "Classification of facebook news feeds and sentiment analysis", 2014

[8] H Margono, X Yi, G Raikundalia, "Mining Indonesian cyber bullying patterns in social networks", Thirty-Seventh Australasian Computer Science Conference, 2014

[9] Scheithauer, Herbert & Petras, Ira-Katharina & Petermann†, Franz, "Cybermobbing Cyberbullying. Kindheit und Entwicklung", 2020

[10] Casas, Jose A. & Ortega-Ruiz, Rosario & Monks, Claire, "Cyberbullying", 2020