CYBER BULLY INDICATOR

Joshi Dhrupad[1], Mayekar Tejas[2], Singh Rajiv ranjan[3], Shinde Sushant[4], Prof. Shweta Patil[5]

[1234]Student, [5]Professor

Department of Computer Engineering

Pillai College Of Engineering, University of Mumbai, New Panvel, India

---

Abstract - A study on cyberbullying data from 2018 through 2020 by Comparitech has shown social media to be a rising medium of bullying. These channels provide few features for dealing with this apart from the basic report feature. Like downvote feature on Reddit, hiding comments on Twitter, new restrict feature on Instagram.However, this feature depicts the action that users can do after the incident. There seems to be no way of dealing with it when the bully is initiating it. We are keen to explore this area and have come up with a solution based on NLP. The approach is to use Big data Analytics on the comments and contents of Instagram to prepare a dataset on which we will train our model. The model would apply NLP approaches to train on the dataset. After achieving satisfactory accuracy this model will be used on the server side. Here. the model will label every comment/message with a label that describes the level of it being cyber bullying. A sudden progression in a series of comments will send a signal to the server which will in turn warn the sender of a permanent account ban if he further continues it.

Keywords - Cyberbullying detection, Natural Language Processing, Sentiment analysis, Deep learning,

1.INTRODUCTION

A study on cyberbullying data from 2018 through 2020 by Comparitech has shown that social media to be a rising medium of bullying. These channels provide few features for dealing with this apart from the basic report feature. Like new restrict feature on Instagram, downvote feature on Reddit, hiding comments on Twitter. However, this feature depicts the action that users can do after the incident. There seems to be no way of dealing with it when the bully is initiating it. In this project we will extract the possible keywords that can raise the flag of cyberbullying and by applying the methods of Natural Language Processing and Machine Learning will be able to restrict the actions that can negatively impact an individual and the society.

2. Literature Survey

After studying a few literature papers we had a brief insight into the topic of interest . Following are the points jotted down that pertain to the studied papers:

A. Opinion Mining and Sentiment Analysis: The first paper that we reviewed[1] described a brief overview of NLP. It then provides the techniques that can be used to perform opinion mining and sentiment analysis. Additionally, it also discusses frequent issues encountered and recommendations for avoiding them.

B. Machine learning Based techniques: The literature provided us with two methods from machine learning techniques. Naive-Bayesian filter and Support Vector Machine. After numerous attempts with different kernels, SVM with poly kernel was found to be most precise. [2]

C. Deep learning Based approach: The machine learning techniques however also concluded the compatibility with a single platform. In order to resolve this problem, deep learning based approaches were reviewed[3]. The successive training of models on these platforms points toward the supremacy of deep learning based approach over machine learning.

D. Automatic detection System: While reviewing the last paper we felt the complexity attributed to a deep learning based system. An in-depth analysis of 22 studies on automatic cyberbullying detection[4] discussed the issues that are faced due to the wrong interpretation of "Cyberbullying" as a term. The studies were directed towards a viewpoint that is more coherent with the definition and representation of the phenomenon, so as to have a practical and impactful application.

2.1 Summary of Related Work

The summary of methods used in existing systems is given in Table 1 and the issues identified in them is given in Table 2.

Literature	Summary
Cyberbullying detection on twitter [5]	● Matches tweet from database and keeps a record (can view bully details like locations and tweets) ● Made using PHP, Mysql, Html

	(Lightweight and easy to implement)
Cyber Bullying Detection System [6]	● ML algorithm to find negative comments of user based on there previous chats (easy to implement with better efficiency) ● Changes negative components of the sentence into positive components (gives a positive feedback to receiver)
Scalable, Timely Detection of Cyberbullying in Online Social Networks[7]	● Consist of Dynamic priority scheduler, novel incremental classifier and an initial predictor (High level algorithms gives higher efficiency rate) ● Detects only cyberbullying on vine using only eight 1 GB AWS VM instances (compared to its reports its compact and requires less size)
Online Social Network Bullying Detection Using Intelligence Techniques [8]	● Uses Fuzzy learning algorithm and Naïve classifier technique (can get the output to higher efficiency) ● Extra words are removed using ML (removes extra, unnecessary words for database that improves efficiency)

Table 1 Summary of existing systems

Literature	Issues
Cyberbullying detection on twitter [5]	● Uses twitter API (restricted to twitter) ● Designed to be used by organization (requires a dedicated person) ● Detects only english language with proper format (can not detect emoticons and punctuations)
Cyber Bullying Detection System [6]	● Tores users ways of chatting in database for analysis and comparison (if user increases and use time increase the storage may rise) ● If sentence is not found to be positive it is re-runed through algorithm (can consume huge memory and time.)
Scalable, Timely Detection of Cyberbullying in Online Social Networks[7]	● Use of many algorithms (makes it difficult to implement) ● System would need 120 units of 32 GB VM instances to fully monitor Instagram-scale traffic (high cost and storage for instagram based detection)
Online Social Network Bullying Detection Using Intelligence Techniques [8]	● Uses data from formsprings.me (restricted to keywords from formsprings.me) ● Designed to only extract cyberbullying keywords from input (give list of keywords cannot stop or detect.)

Table 2 Issues identified in existing systems

3. Proposed Work

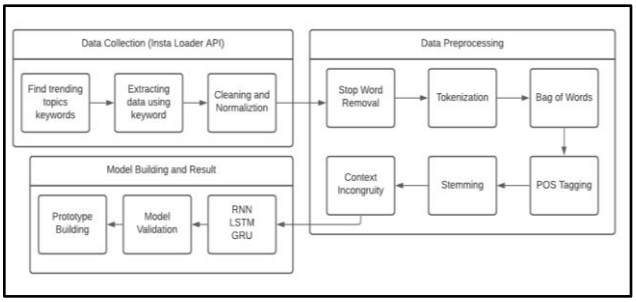

Fig 3.1 Classification of domain techniques

This system enables us to keep track of keywords related to cyber bullying based on previous posted content by various users. This can be primarily done by following three major steps :

A. Data Collection: The initial step in the project is collection of data. We need to ensure that the data collection isn't biased to just one type of data or keywords. We will be extracting the trending keywords known as 'hashtags' by using web scraping tools and later feed those keywords to the 'Insta Loader API'

that will extract the comments and posts from instagram.

B. Data Processing: Comments and post extracted may not be desired form, hence there is a need for data preprocessing. This step will make sure that the data will make some sense that can be later used for detecting the cyber bullying post

C. Model Building and Result: Based on the extracted and preprocessed data , the machine learning model will be built which can detect the bullying comment and post. The results or the accuracy outcome can be calculated and can be measured for its true positive and true negative results.

3.1 System Architecture

The system architecture is given in Figure 1.

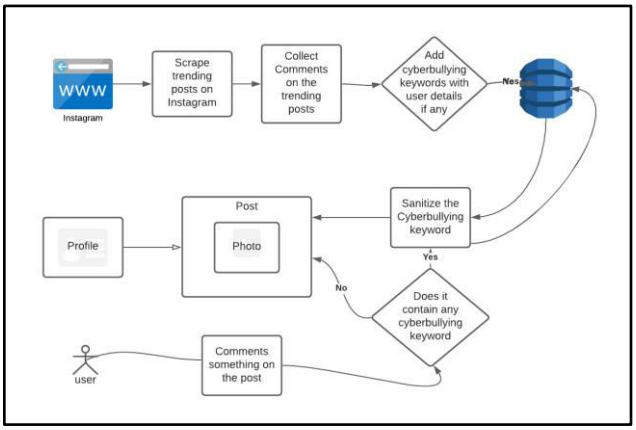

Fig. 3.1.1 Proposed system architecture

In this first the system will scrape the trending topics on instagram,then it will collect all the comments posted by the users on that topics .Then the algorithm will process those comments and give out cyberbullying keywords along with the username and will be stored in a database.
Next when some user will post a comment on someone's post it will take that comment and compare words with the system database if any cyberbullying keyword is detected the world will be striked out and the username and word will be added to the database,the user will receive a warning.

3.2 Implementation Details

In order to obtain the required results we will be following the mentioned techniques.

A. Algorithm/Technique

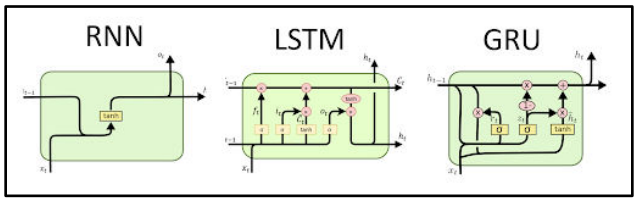

Fig 3.2.1 Models to be used in proposed system

LSTM (Long short-term memory) :
It belongs to the class of recurrent neural networks architecture. Due to its feedback connection, it is highly used in deep learning. It is able to process signals as simple as images to as complex as speech and video. Each unit of LSTM consists of a cell, an input gate, an output gate and a forget gate. It's applications include handwriting recognition, speech recognition, anomaly detection in network traffic and much more.

GRU (Gated recurrent unit):
It is another variation of recurrent neural network architecture. It is quite similar to LSTM as it is LSTM with a forget gate. Its unit consists of an Update gate, Reset gate and current memory gate. It doesn't use an internal cell state. It is quite useful for smaller datasets.

B. Use Case Diagram / Activity Diagram
Given below is an activity diagram of the proposed system.

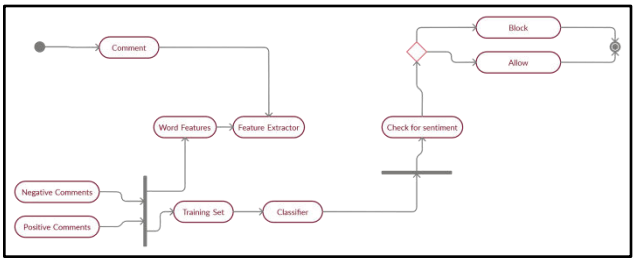

Fig 3.2.2 Activity diagram of the proposed system

When a user comments on any post it will be extracted from the text field and feature engineering will be performed which is built on pre existing data comments. This extracted data will be fed on the trained system to classify users sentiments and later judgement will be performed whether to allow or restrict that comment or sanitize and post that particular comment.

3.3 Dataset and Parameters

The sample dataset used in the experiment is extracted using an API which is in raw JSON file format. Sample data given in figure 3.1

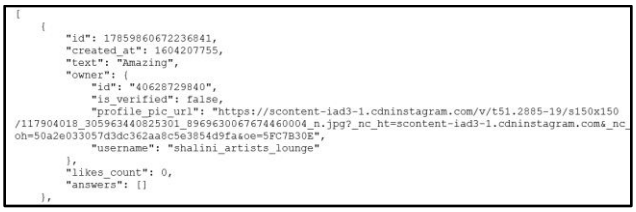

Fig 3.3.1 Sample Dataset Used for Experiment

A. Evaluation Metrics

Below mentioned metrics are used to evaluate the performance of trained models.

Cosine Similarity:

It talks about the similarity between two non-zero vectors. It is given by the below equation whose result lies between 0 and 1:

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

In our case the words are converted into non- zero vectors by using mining algorithms such as TF-IDF or Bag Of Words as an instance.

Perplexity:

It is a term that measures the accuracy of a probability model. It is given by the below equation:

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}}$$

In our case it is computed per word and relies on the underlying probability distribution of the words in the sentences.

ACKNOWLEDGMENT

We would like to express our specials thanks to Principal Dr. Sandeep Joshi for sharing his pearls of wisdom with us during this course of this project. We pay our deep sense of gratitude to the Head Of the Computer Science Department, Dr Sharvari Govilkar to encourage us to the highest peak and to provide us the opportunity to prepare the project. We acknowledge our project guide Prof. Shweta Patil whose valuable guidance and kind supervision given throughout the course which shaped the present work as it shows and giving apt guidance for this project.

REFERENCES

1. Noviantho, Sani Muhamad Isa, Livia Ashianti. Cyberbullying Classification using Text Mining, 2017.
2. Yasir Ali, Solangi, Zulfiqar Ali Solangi, Samreen Aarain1, Amna Abro, Ghulam Ali Mallah, Asadullah Shah. Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis, 2018.
3. Maral Dadvar and Kai Eckert. Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study, 2018.
4. H. Rosa, N. Pereira, R. Ribeiro, P.C. Ferreiraa, J.P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A.M. Veiga Simão, I. Trancoso. Automatic cyberbullying detection: A systematic review, 2019.
5. Liew Choong Hon, Kasturi Dewi Varathan ,"Cyberbullying Detection System on Twitter",Vol 1 No.1 International journal of Information Systems and Engineering , April 2015.
6. Giovanni Berrios ,Chanhee Shin ,Nishal Kallupalle ,"Cyber Bullying Detection System", Senior Design 4739 Professor Ashis, 5th June 2020.
7. B.Sri Nandhinia, J.I.Sheeba, "Online Social Network Bullying Detection Using Intelligence Techniques," ScienceDirect.Procedia Computer Science 45 (2015) 485 – 492 ,International Conference on Advanced Computing Technologies and Applications, 2015.
8. Rahat Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra,"Scalable, Timely Detection of Cyberbullying in Online Social Networks", 2018.