Detection of Cyberbullying on Social Media Using Machine learning

Sadineni Yaswanth Sai[1], Yarramuri Pavan Kumar[2], Shaik Salman[3], Prashant Singamsetti[4]

[1] *UG Student, Dept of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India*
[2] *UG Student, Dept of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India*
[3] *UG Student, Dept of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India*
[4] *UG Student, Dept of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India*

ABSTRACT

Cyberbullying represents a significant challenge in the online realm, impacting both teenagers and adults and giving rise to severe consequences such as suicide and depression. The imperative for regulating content on social media platforms has become increasingly evident. In response to the incidents of harm caused by cyberbullying, we employ data sourced from two distinct manifestations of this phenomenon: hate speech tweets on Twitter. Our objective is to construct a model utilizing the Support Vector Machine (SVM) algorithm in the field of machine learning to detect instances of cyberbullying within textual data. The pervasive nature of cyberbullying has necessitated a proactive approach to mitigate its adverse effects on individuals, particularly the vulnerable demographic of teenagers. The alarming rise in incidents resulting in tragic outcomes like suicide and depression underscores the urgency of addressing this issue. Recognizing the role of social media platforms as conduits for cyberbullying, there is a compelling need for effective content regulation to create a safer online environment. To tackle this challenge, we leverage data derived from hate speech tweets on the Twitter platform, representing two distinct forms of cyberbullying. Employing the Support Vector Machine (SVM) algorithm within the realm of machine learning, we aim to develop a robust model capable of identifying cyberbullying instances embedded in textual data. This approach reflects a proactive stance in combating the detrimental impacts of cyberbullying, emphasizing the role of technological interventions and algorithmic solutions in fostering a more secure and supportive online space. In addition, continuous monitoring and adaptation of the model will be crucial to staying ahead of evolving cyberbullying tactics and safeguarding the well-being of internet users.

Keyword : - *Cyberbullying, K-Nearest Neighbor, Support Vector Machine, Machine Learning*

1. INTRODUCTION

In the era of ubiquitous online interactions, the prevalence of cyberbullying poses a serious threat to the well-being of individuals, especially adolescents. This research project endeavors to address this pressing societal concern by employing advanced machine learning techniques and feature extraction methods for the detection of cyberbullying in textual data. By meticulously preprocessing and refining the raw text through techniques such as replacing contractions and stopword removal, we aim to enhance the effectiveness of subsequent analysis. 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].

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 cyber bully 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. The integration of Support Vector Machine

(SVM) and k-Nearest Neighbors (KNN) algorithms contributes to a robust and multifaceted approach, seeking to mitigate the impact of cyberbullying and create a safer online environment.

2. EXISTING SYSTEM

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 topicsensitive classifiers are more effective to detect cyberbullying. 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. 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. Among the different classification techniques, SVM gets notable attention due to better performance in various text classifications. Hence the aim of this research was to explore various machine learning algorithms.

3. PROPOSED METHODOLOGY

Our project aims to detect major form of cyberbullying on Twitter by detecting hate speech and classifying them as containing cyberbullying or not. Below figure Fig 1 describes the process of the project.

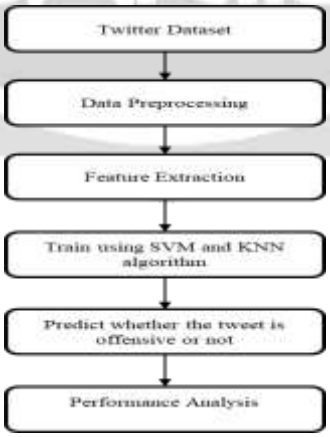

Fig -1: Flow Chart

3.1 Dataset

A total of 31962 individual data points are included in the dataset. 65 percent of the dataset is offensive, while 35 percent is non-offensive. In the total dataset 70 percent is used for training of data and 30 percent of data for testing. The below figure shows some examples in the dataset. The datset consists of three columns id, label and tweet.

id	label	tweet
1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disappointed #getthanked
3	0	!ohday your majesty
4	0	#model i love u take with u all the time in urÃ°Ã¼Ã¤Ã«!!! Ã°Ã¼Ã¤Ã¬Ã°Ã¼Ã¤Ã¼Ã°Ã¼Ã¤Ã¯Ã°Ã¼Ã¤Ã® Ã°Ã¼Ã¤Ã°Ã°Ã¼Ã¤Ã±Ã°Ã¼Ã¤Ã²
5	0	factsguide: society now #motivation
6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo
7	0	@user camping tomorrow @user @user @user @user @user @user @user dannyÃ¢Â€Â¦
8	0	the next school year is the year for exams.Ã°Ã¼Ã˜Ã¢ can't think about that Ã°Ã¼Ã˜Ã¢ #school #exams #hate #imagine #actorslife #revolutionschool #girl
9	0	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers Ã¢Â€Â¦
10	0	@user @user welcome here ! i'm it's so #gr8 !
11	0	Ã¢Â€Âº #ireland consumer price index (mom) climbed from previous 0.2% to 0.5% in may #blog #silver #gold #forex
12	0	we are so selfish. #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproblems #selfish #heartbreaking #values #love #
13	0	i get to see my daddy today!! #80days #gettingfed
14	1	@user #cnn calls #michigan middle school 'build the wall' chant." #tcot
15	1	no comment! in #australia #opkillingbay #seashepherd #helpcovedolphins #thecove #helpcovedolphins
16	0	ouch...junior is angryÃ°Ã¼Ã˜Ã¡got7 #junior #yugyoem #omg
17	0	i am thankful for having a paner. #thankful #positive
18	1	retweet if you agree!
19	0	its #friday! Ã°Ã¼Ã˜Ã smiles all around via ig user: @user #cookies make people
20	0	as we all know, essential oils are not made of chemicals.
21	0	#euro2016 people blaming ha for conceded goal was it fat rooney who gave away free kick knowing bale can hit them from there.
22	0	sad little dude.. #badday #coneofshame #cats #pissed #funny #laughs
23	0	product of the day: happy man #wine tool who's it's the #weekend? time to open up & drink up!
24	1	@user @user lumpy says i am a . prove it lumpy.
26	0	@user #tgif #ff to my #gamedev #indiedev #indiegamedev #squad! @user @user @user @user @user
27	0	beautiful sign by vendor 80 for $45.00!! #upsideofflorida #shopalyssas #love

Fig -2: Dataset

3.2 Data Preprocessing

In the process of preparing the Twitter hate speech dataset for analysis or machine learning tasks, several crucial data preprocessing steps have been implemented. The initial stage involves the removal of web links (URLs) from the text data, where a regular expression is utilized to identify and replace URLs with an empty string. This step aims to eliminate hyperlinks that might not contribute meaningfully to the analysis, ensuring a more focused and relevant dataset. Following this, the focus shifts to removing punctuation, targeting symbols such as commas, periods, and hashtags. A translation table is constructed to map each punctuation character to an empty space, simplifying the text and reducing noise in the data. Subsequently, a more comprehensive data cleaning process is executed, involving the removal of the 'id' column from the dataset. Additionally, rows containing missing values are dropped to enhance data quality and consistency, creating a cleaner dataset for subsequent analysis. The emphasis on data cleanliness is pivotal in mitigating potential biases or inaccuracies in the analysis, fostering a more reliable foundation for drawing meaningful conclusions. Furthermore, the data is strategically split into two components: features and labels. The 'tweet' column, serving as the feature, is extracted to represent the input data for the subsequent analysis or machine learning tasks. The utilization of the train-test split technique ensures the separation of data for training and evaluation, contributing to the robustness of the model's performance. These preprocessing techniques collectively contribute to the refinement of the Twitter hate speech dataset, addressing issues such as irrelevant information, missing values, and data organization. Overall, the meticulous execution of these steps lays the groundwork for a more accurate and insightful exploration of cyberbullying and hate speech trends on social media platforms.

3.3 Feature Extraction

Feature extraction plays a crucial as text data needs to be converted into numerical form for classification by classifiers. This project explores three feature extraction methods: replacing contractions, regex tokenization, and stopword removal. Additionally, Porter Stemmer is applied to further refine the text data. The research investigates three widely-used feature extraction models: Bag of Words (BoW), TF-IDF, and Word2Vec. BoW is a simple

method that focuses on the occurrence of words within a document, using a vocabulary derived from all documents. TF-IDF, similar to BoW, creates a vocabulary and addresses the frequency of words in a document compared to the entire corpus. It introduces term frequency and inverse document frequency components. Word2Vec, a neural network-based method, employs word embeddings to represent words in vector form, facilitating similarity calculations between words. The Word2Vec model includes Common Bag of Words (CBOW) and Skip Gram methods for constructing word embeddings. CBOW predicts a word based on multiple context words, while Skip Gram predicts multiple context words using a single input word. Both methods use forward and backpropagation to train neural networks and find optimal parameters. The resulting feature vectors are created by concatenating and combining word vectors in a document, either through summation or averaging. Each feature extraction method addresses specific challenges, with BoW being simple but effective for sentiment analysis, TF-IDF addressing issues in BoW related to word frequency, and Word2Vec incorporating word embeddings for nuanced semantic understanding.

3.4 Model Training

Using the training data the following classifiers will be trained and tested on:

- Support Vector Machine: A powerful supervised learning algorithm, is employed to discern patterns and relationships within the feature-extracted data, facilitating the creation of a robust model for cyberbullying detection. Its ability to handle high-dimensional data and nonlinear relationships proves valuable in capturing intricate nuances present in the textual content, contributing to enhanced model performance.
- KNN: Enriches the model's capacity to classify instances of cyberbullying based on the proximity of data points in the feature space. KNN, being a non-parametric and instance-based algorithm, leverages the similarity between data points to make predictions. Its adaptability to varying data distributions makes it a valuable asset in capturing local patterns, which might be particularly beneficial in the context of cyberbullying detection where the nature of offensive language and context can be diverse.

3.4 Performance Analysis

True Positives(TP), True Negatives(TN), False Postitives (FP), Flase Negatives(FN)

- Accuracy(A): Accuracy is a metric that measures how often a machine learning model correctly predicts the outcome. You can calculate accuracy by dividing the number of correct predictions by the total number of predictions.

$$\text{Accuracy(A)} = (TP + TN) / (TP + TN + FP + FN)$$

- Precision(P): The percentage of correctly predicted positive outcomes out of all the predicted positive outcomes.

$$\text{Precision} = TP / (TP+FP)$$

- Recall is a metric that measures how often a machine learning model correctly identifies positive instances (true positives) from all the actual positive samples in the dataset.

$$\text{Recall} = TP / (TP+FN)$$

4. CONCLUSIONS

In conclusion, this research project employs advanced feature extraction techniques and machine learning algorithms, including Support Vector Machine (SVM) and k-Nearest Neighbors (KNN), to tackle the pervasive issue of cyberbullying. Through meticulous data preprocessing and the exploration of diverse feature extraction models, such as Bag of Words, TF-IDF, and Word2Vec, the study demonstrates a comprehensive approach to capturing nuanced textual patterns. The integration of SVM and KNN showcases a synergistic strategy, leveraging SVM's strength in handling complex relationships and KNN's adaptability to local patterns. This research contributes valuable insights to the evolving field of cyberbullying detection, offering a multifaceted model that addresses the diverse challenges posed by online harassment and fosters a safer digital space. This multifaceted strategy not only enhances the model's accuracy but also provides valuable insights into the diverse manifestations of cyberbullying. The knowledge gained from this research contributes to the ongoing efforts to create not just a technologically robust but also a socially responsible online environment, mitigating the impact of cyberbullying and fostering a culture of empathy and inclusivity.

	SVM	KNN
Accuracy	0.96	0.93
Precision: 0	0.96	0.94
: 1	0.90	0.99
Recall: 0	1.0	1.0
: 1	0.5	0.13

Table -1 : Results

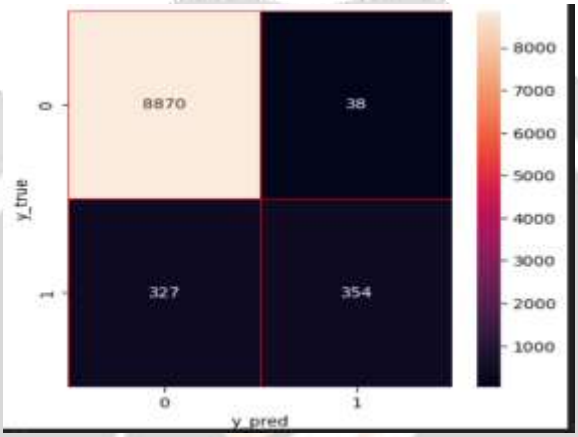

Fig -3: Confusion Matrix(SVM)

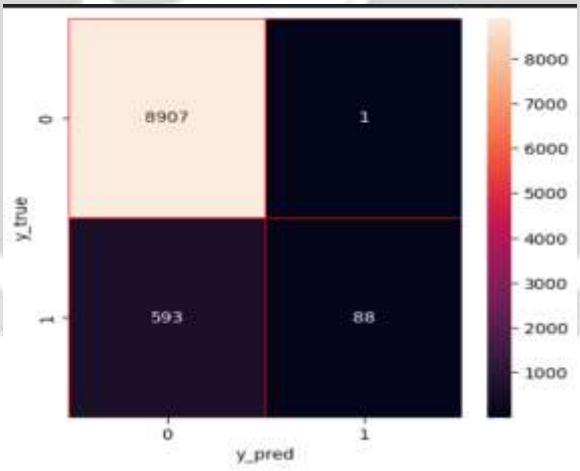

Fig -4: Confusion Matrix(KNN)

6. REFERENCES

[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-Bangladeshjanuary-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]. A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," 2015, doi: 10.1109/EIT.2015.7293405.
[5]. R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," 2016, doi: 10.1145/2833312.2849567.
[6]. V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network," 2019, doi: 10.1109/ICACCS.2019.8728378.
[7]. A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," Artif. Intell. Rev., vol. 53, no. 6, 2020, doi: 10.1007/s10462-019-09794-5.
[8]. 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.
[9]. S. Hnduja and J. W. Patchin "Cyberbullying: Identification, Prevention, & Response,"Cyberbullying Res. Cent, no. October, pp. 1-9,2018.
[10]. 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.
[11]. 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.
[12]. 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.