Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**

Journal of Nonlinear Analysis and Optimization : Theory & Applications

Editors-in-Chief :
Sompong Dhompongsa
Somyot Plubtieng

A Comparative analysis of Machine Learning Models for Detection of Cyberbullying and Abuser Profile on Social Media

HIMAM BASHA SHAIK[1], SD.HABI[2]

[1]Assistant Professor, Dept. of MCA, QIS College of Engineering and Technology, Ongole, Andhra Pradesh.

[2]PG Scholar, Dept. of MCA, QIS College of Engineering and Technology, Ongole, Andhra Pradesh.

ABSTRACT— In today's digital era, the escalating phenomenon of cyberbullying is a pervasive and growing concern. With the increasing prevalence of social media platforms, such as Twitter, online abusive behavior has become a significant issue that often leads to unpleasant experiences for users. Manual detection of abnormal and bullying behavior within the realm of social media is inherently not scalable. Moreover, most existing studies on cyberbullying detection have been predominantly conducted in English and very limited work has been done on Urdu(awidelyusedlanguagein Asia). This paper presents an approach for detecting cyberbullying in Roman Urdu tweets and identifying abuser profiles on Twitter. Firstly, we develop a text corpus of Roman Urdu tweets with user profile data. Subsequently, we employ Gated Recurrent Unit (GRU) model coupled with the application of word2vec technique for word embedding to develop a cyberbullying detection model. Furthermore, we present temporal abusive tweet probability analysis method to provide anuanced analysis of thenumberofbullyingandnon-bullyingtweetssentbyindividualswithinaspecifictime interval. To evaluate the performance, we compare the GRU-based approach with other machine learning models.

Index Terms – Cyberbullying, multi-modality, social media, hierarchy attention.

I. INTRODUCTION

Social media networks such as Facebook, Twitter, Flickr, and Instagram have become the preferred online platforms for interaction and socialization among people of all ages. While these platforms enable people to communicate and interact in previously un thinkable ways, they have also led to malevolent activities such as cyber-bullying. Cyber bullying is a type of psychological abuse with a significant impact on society. Cyber-bullying events have been increasing mostly among young people spending most of their time navigating between different social media platforms. Particularly, social media networks such as Twitter and Facebook are prone to CB because of their popularity and the anonymity that the Internet provides to abusers. In India, for example, 14 percent of all harassment occurs on Facebook and Twitter, with 37 percent of these incidents involving youngsters.

Moreover, cyber bullying might lead to serious mental issues and adverse mental health effects. Most suicides are due to the anxiety, depression, stress, and social and emotional difficulties from cyber-bullying events. This motivates the need for an approach to identify cyber bullying in social media messages (e.g., posts, tweets, and comments). In this article, we mainly focus on the problem of cyber bullying detection on the Twitter platform. As cyber bullying is becoming a prevalent problem in Twitter, the detection of cyber bullying events from tweets and provisioning preventive measures are the primary tasks in battling cyber bullying threats. Therefore, there is a greater need to increase the research on social networks-based CB in order to get greater in sights and aid in the development of effective tools and approaches to effectively combat cyber bullying problem. Manually monitoring and controlling cyber bullying on Twitter platform is virtually impossible. Furthermore, mining social media messages for cyber bullying detection is quite difficult. For example, Twitter messages are often brief, full of slang, and may include emojis, and gifs, which makes it impossible to deduce individuals' intentions and meanings purely from social media messages. Moreover, bullying can be difficult to detect if the bully uses strategies like sarcasm or passive-aggressiveness to conceal it.

Despite the challenges that social media messages bring, cyber bullying detection on social media is an open and active research topic. Cyber bullying detection within the Twitter platform has largely been pursued through tweet classification and to a certain extent with topic modeling approaches. Text classification based on supervised machine learning (ML) models are commonly used for classifying tweets into bullying and non-bullying tweets

II. LITERATURE SURVEY

A. *Cyber bullying Detection and Hate Speech Identification using Machine Learning Techniques*

Bullying has been prevalent since the beginning of time, It’s just the ways of bullying that have changed over the years, from physical bullying to cyberbullying. According to Williard (2004), there are eight types of cyberbullying such as harassment, denigration, impersonation, etc. It’s been around 2 decades since social media sites came into the picture, but there haven’t been a lot of effective measures to curb social bullying and it has become one of the alarming issues in recent times.Our paper presents an analytical review of cyberbullying detection approaches and assesses methods to recognize hate speech on social media. We aim to apply traditional supervised classification methods as well as some novel ensemble machine learning techniques using a manually annotated open-source dataset for this purpose. This paper does a comparative study of various Supervised algorithms, including standard, as well as ensemble methods. The evaluations of the result based upon the scores obtained by accuracy shows that Ensemble supervised methods have the potential to perform better than traditional supervised methods.

B. *Cyber bullying detection on social media using machine learning*

Usage of internet and social media backgrounds tends in the use of sending, receiving and posting of negative, harmful, false or mean content about another individual which thus means Cyberbullying. Bullying over social media also works the same as threatening, calumny, and chastising the individual. Cyberbullying has led to a severe increase in mental health problems, especially among the young generation. It has resulted in lower self-esteem, increased suicidal ideation. Unless some measure against cyberbullying is taken, self-esteem and mental health issues will affect an entire generation of young adults. Many of the traditional machine learning models have been implemented in the past for the automatic detection of cyberbullying on social media. But these models have not considered all the necessary features that can be used to identify or classify a statement or post as bullying. In this paper, we proposed a model based on various features that should be considered while detecting cyberbullying and implement a few features with the help of a bidirectional deep learning model called BERT.

C. Aggression detection in social media from textual data using deep learning models

With the advancement of technology, social media such as Facebook, Twitter, etc. plays an important role in communication whether it is texting, sharing photos, audio-video calls or expressing views through comments. Along with these advantages, it has some negative sides as well which brings aggression towards some section of people. Such aggression, hatred in social media needs to be detected and prevented automatically which is the main objective of our work. We have worked on Hindi, English and Hindi-English (code-mixed) datasets. We used features like word vectors, aggressive words (manually created dictionary), sentiment scores, parts of speech and emojis for the classification task. We experimented with several machine learning and deep learning models and the results indicate that XGBoost Classifier, Gradient Boosting Classifier (GBM) and Support Vector Machine (SVM) are most suited for the task. Therefore the output of the three classifiers were used for majority voting which provides f-scores of 68.13, 54.82 and 55.31 for the English, Hindi and code-mixed datasets respectively.

III. PROPOSED SYSTEM

The overview of our proposed system is shown in the below figure.

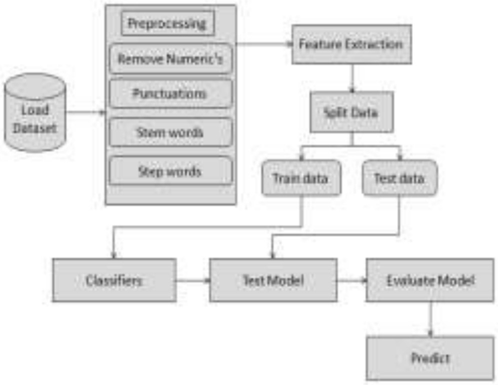


Fig. 1: System Overview

Implementation Modules

Service Provider Module

- ✓ In this module, service provider login to the system using valid username and password. After login successful, he can perform the following operations like train and test dataset, view trained and tested accuracy and view remote users.

Train and Test Model

- ✓ In this module, the service provider split the Used dataset into train and test data of ratio 70 % and 30 % respectively. The 70% of the data is consider as train data which is used to train the model and

30% of the data is consider as test which is used to test the model

Remote User

- ✓ In this module, the remote user register to the system, and login to the system valid username, and password. After login successful, he can perform view profile, Identifying abuser profile.

Graphical Analysis

- ✓ In this module, display the graphs like accuracy and predicted ratio of the system. Various factors take into consideration for the graph analysis. In this phase plot the charts like bar chart and so others.

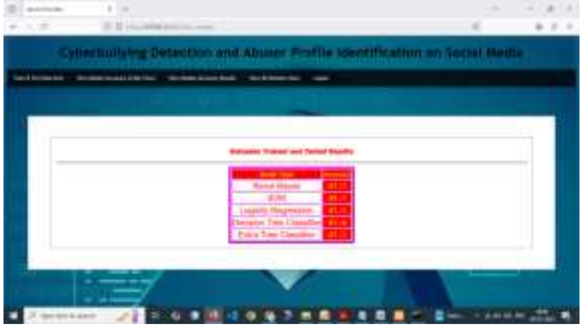
IV. RESULTS

Cyberbullying Detection and Abuser Profile Identification on Social Media

Fig.2: Home Page

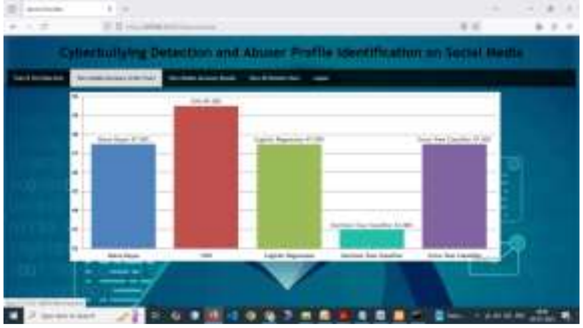
Cyberbullying Detection and Abuser Profile Identification on Social Media

Fig.3: Service Provider Login



Cyberbullying Detection and Abuser Profile Identification on Social Media

Fig.4: Model Accuracy



Cyberbullying Detection and Abuser Profile Identification on Social Media

Fig.5: Model Accuracy Results

Fig.6: Prediction

V. CONCLUSION

The findings of this study underscore the critical challenges faced by of cyberbullying content detection, primarily stemming from the limited availability of comprehensive datasets containing user profile information. This scarcity imparts significant implications for the accurate identification and categorization of abusive content spanning diverse cyberbullying categories in the Roman Urdu text on social media. The proposed method, utilizing the GRU model with pre-learned embeddings, exhibited remarkable advancements over conventional machine learning and deep learning models, including LSTM, Bi-LSTM with attention layers, and CNN. These advancements were observed across key performance metrics such as accuracy, precision, and F-measure, with the GRU-based method attaining an accuracy rate of 97%, outperforming alternative methods in this domain. Furthermore, our investigation highlighted the effectiveness of employing Decision Tree and Gradient Boosting as conventional machine learning classifiers in conjunction with TF-IDF for cyberbullying detection. These classifiers demonstrated relatively superior performance, potentially attributed to the incorporation of lexical nor malization during the data pre-processing phase. This stan dardization of Roman Urdu words contributed to improved performance and accuracy. The proposed method exhibited the capability to successfully classify users into three distinct categories based on their cyberbullying behavior: Normal Users (922 instances), Suspected Users (57 instances), and Abusive Users (25 instances).

REFERENCES

[1] F. Mishna, M. Khoury-Kassabri, T. Gadalla, and J. Daciuk, ``Risk factorsfor involvement in cyber bullying: Victims, bullies and bullyvictims," Children Youth Services Rev., vol. 34, no. 1, pp. 6370, Jan. 2012, doi: 10.1016/j.childyouth.2011.08.032.

[2] K. Miller, ``Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike, and the law's limited available redress," Southern California Interdiscipl. Law J., vol. 26, no. 2, p. 379, 2016.

[3] A. M. Vivolo-Kantor, B. N. Martell, K. M. Holland, and R. Westby, ``A systematic review and content analysis of bullying and cyber-bullying measurement strategies," Aggression Violent Behav., vol. 19, no. 4, pp. 423434, Jul. 2014, doi: 10.1016/j.avb.2014.06.008.

[4] H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, ``Associations between cyberbullying and school bullying victimization and suicidal ideation, plan and attempts among Canadian schoolchildren," PLoS ONE, vol. 9, no. 7, Jul. 2014, Art. no. e102145, doi: 10.1371/journal.pone.0102145.

[5] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, ``Improving cyberbullying detection with user context," in Proc. Eur. Conf. Inf. Retr., in Lecture Notes in Computer Science: Including Subseries Lecture Notesin Articial Intelligence and Lecture Notes in Bioinformatics, vol. 7814,2013, pp. 693696.

[6] A. S. Srinath, H. Johnson, G. G. Dagher, and M. Long, ``BullyNet: Unmasking cyberbullies on social networks," IEEE Trans. Computat. Social Syst., vol. 8, no. 2, pp. 332344, Apr. 2021, doi: 10.1109/TCSS.2021.3049232.

[7] A. Agarwal, A. S. Chivukula, M. H. Bhuyan, T. Jan, B. Narayan, and M. Prasad, ``Identication and classication of cyberbullying posts: A recurrent neural network approach using under-sampling and class weighting," in Neural Information Processing (Communications in Computer and Information Science), vol. 1333, H. Yang, K. Pasupa, A. C.-S. Leung, J. T. Kwok, J. H. Chan, and I. King, Eds. Cham, Switzerland: Springer, 2020, pp. 113120.

[8] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, ``Machine learning and feature engineering-based study into sarcasm and irony classication with application to cyberbullying detection," Inf. Process. Manage., vol. 58, no. 4, Jul. 2021, Art. no. 102600, doi: 10.1016/j.ipm.2021.102600.

[9] N. Yuvaraj, K. Srihari, G. Dhiman, K. Somasundaram, A. Sharma, S. Rajeskannan, M. Soni, G. S. Gaba, M. A. AlZain, and M. Masud, ``Nature-inspired-based approach for automated cyberbullying classication on multimedia social networking," Math. Problems Eng., vol. 2021, pp. 112, Feb. 2021, doi: 10.1155/2021/6644652.

AUTHORS Profile

Mr. Himam Basha Shaik is an Assistant Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He earned his Master of Computer Applications (MCA) from Anna University, Chennai. With a strong research background, He has authored and co-authored research papers published in reputed peer-reviewed journals. His research interests include Machine Learning, Artificial Intelligence, Cloud Computing, and Programming Languages. He is committed to advancing research and fostering innovation while mentoring students to excel in both academic and professional pursuits.

Mr. SD. Habi has received his B.Sc (Computers) degree from ANU 2022 and Pursuing MCA QIS College of Engineering And Technology Affiliated to JNTUK 2023-2025.