International Journal of Multidisciplinary Engineering in Current Research - IJMEC
Volume 10, Issue 5, May-2025, http://ijmec.com/, ISSN: 2456-4265

Dea-Rnn A Hybrid Deep Learning Approach For Cyberbullying Detection In Twitter Social Media Platform

Vattam Lokesh
PG scholar, Department of MCA, CDNR collage, Bhimavaram, Andhra Pradesh.
K.Sridevi
(Assistant Professor), Master of Computer Applications, DNR collage, Bhimavaram, Andhra Pradesh.

Abstract: Cyberbullying on social media has become a significant concern due to the widespread use of online platforms. Unlike traditional bullying, cyberbullying can occur anonymously, instantly, and on a massive scale, making it harder to detect and control. This paper proposes an AI-based detection system that identifies offensive, threatening, or harmful comments on social media. Using natural language processing (NLP) techniques along with machine learning and deep learning models, the system is trained to distinguish between normal and abusive language patterns. The model is evaluated using labeled datasets, demonstrating high accuracy in detecting cyberbullying across various social media texts. This approach aims to assist platform moderators and create safer digital environments.

I. INTRODUCTION

Now more than ever technology has become an integral part of our life. With the evolution of the internet. Social media is trending these days. But as all the other things mis users will pop out sometimes late sometime early but there will be for sure. Now Cyber bullying is common these days.

Sites for social networking are excellent tools for communication within individuals. Use of social networking has become widespread over the years, though, in general people find immoral and unethical ways of negative stuff. We see this happening between teens or sometimes between young adults. One of the negative stuffs they do is bullying each other over the internet. In online environment we cannot easily said that whether someone is saying something just for fun or there may be other intention of him. Often, with just a joke, "or don't take it so seriously," they'll laugh it off Cyber bullying is the use of technology to harass, threaten, embarrass, or target another person. Often this internet fight results into real life threats for some individual. Some people have turned to suicide. It is necessary to stop such activities at the beginning. Any actions could be taken to avoid this for example if an individual's tweet/post is found offensive then maybe his/her account can be terminated or suspended for a particular period. So, what is cyber bullying??

Cyber bullying is harassment, threatening, embarrassing or targeting someone for the purpose of having fun or even by well-planned means

II. LITERATURE SURVEY

- **Dinakar et al. (2011)** focused on classifying YouTube comments to detect cyberbullying across topics like sexuality, race, and intelligence. Their study used Support Vector Machines (SVM) and highlighted the need for topic-sensitive models due to the diversity of harmful content. They emphasized that context-aware filtering can significantly improve bullying detection rates.

- **Xu et al. (2012)** presented a framework using text classification to detect bullying language on Twitter. By applying n-gram features and sentiment analysis, their system differentiated between aggressive and non-aggressive tweets. The results showed that sentiment combined with textual patterns can effectively enhance classification performance.

- **Dadvar et al. (2013)** improved cyberbullying detection by incorporating user-specific features like age, gender, and past behavior. Using a combination of content-based and user-based features, their model showed that integrating user

ISSN: 2456-4265
IJMEC 2025

profiling can increase detection accuracy, especially for frequent offenders.

 Zhao et al. (2016) proposed a deep learning-based approach using Convolutional Neural Networks (CNNs) for cyberbullying detection. Unlike traditional feature engineering, their model learned patterns directly from raw text data. This approach showed significant performance gains in detecting nuanced and disguised bullying.

 Rosa et al. (2019) introduced a multi-language framework using LSTM models to identify bullying across different cultural contexts. Their study demonstrated that deep learning models can be trained to recognize bullying in various languages with minimal manual intervention.

 Agrawal and Awekar (2018) explored the use of Recurrent Neural Networks (RNNs) for cyberbullying detection. They compared multiple neural architectures and concluded that Bi-LSTM (Bidirectional LSTM) networks achieved the best performance due to their ability to capture forward and backward context in abusive language.

 Chavan and Shylaja (2015) used NLP and machine learning to classify Facebook comments as bullying or non-bullying. Their work focused on preprocessing techniques such as stemming and stop-word removal, and they found that Naive Bayes classifiers gave satisfactory results on balanced datasets.

III. PROPOSED METHOD

The proposed methodology aims to detect cyberbullying on social media by leveraging a deep learning model trained on labeled text data. The first step involves **data collection** from public datasets like Kaggle's cyberbullying datasets, which contain comments labeled as bullying, threat, insult, or neutral. The **preprocessing phase** includes removing stop words, special characters, converting text to lowercase, and applying tokenization.

Once preprocessed, the text is converted into **numerical format** using word embeddings like **Word2Vec** or **TF-IDF**. These embeddings are then fed into a **deep learning model**, such as a **Bidirectional LSTM (BiLSTM)** or **CNN**, which is trained to detect sequential and contextual abuse in text. The architecture includes embedding layers, dropout for regularization, and a softmax output layer to classify the comments.

The model is trained on 80% of the data and tested on the remaining 20% to evaluate its performance using metrics such as **accuracy, precision, recall, and F1-score**. Comparative analysis is also done with traditional ML models like SVM and Random Forest to demonstrate the superiority of deep learning in detecting complex patterns of online abuse.

V. CONCLUSION

Cyber bullying across internet is dangerous and leads to mis happenings like suicides, depression etc and therefore there is a need to control its spread. Therefore cyber bullying detection is vital on social media platforms. With availability of more data and better classified user information for various other forms of cyber attacks Cyber bullying detection can be used on social media websites to ban users trying to take part in such activity In this paper we proposed an architecture for detection of cyber bullying to combat the situation. We discussed the architecture for two types of data: Hate speech Data on Twitter and Personal attacks on Wikipedia. For Hate speech Natural Language Processing techniques proved effective with accuracies of over 90 percent using basic Machine learning algorithms because tweets containing Hate speech consisted of profanity which made it easily detectable. Due to this it gives better results with BOW and TF-IDF models rather than Word2Vec models However, Personal attacks were difficult to detect through the same model because the comments generally did not use any common sentiment that could be learned however the three feature selection methods performed similarly.Word2Vec models that use context of features proved effective in both datasets giving similar results in comparatively less features when combined with Multi Layered Perceptrons.

REFERENCES

[1] I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T.
Bermudez, "Towards the detection of cyberbullying based on
social network mining techniques," in Proceedings of 4th
International Conference on Behavioral, Economic, and Socio-
Cultural Computing, BESC 2017, 2017, vol. 2018-January, doi:
10.1109/BESC.2017.8256403.
[2] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and
P. G. Bringas, "Supervised machine learning for the detection of
troll profiles in twitter social network: Application to a real case
of cyberbullying," 2014, doi: 10.1007/978-3-319-01854-6_43.
[3] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative
detection of cyberbullying behavior in Twitter data," 2015, doi:
10.1109/EIT.2015.7293405.
[4] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of
cyberbullying on social networks based on bullying features,"
2016, doi: 10.1145/2833312.2849567.
[5] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection
of Cyberbullying Using Deep Neural Network," 2019, doi:
10.1109/ICACCS.2019.8728378.
[6] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine
learning to detect cyberbullying," 2011, doi:
10.1109/ICMLA.2011.152.
[7] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection
using Pre-Trained BERT Model," 2020, doi:
10.1109/ICESC48915.2020.9155700.
[8] M. Dadvar and K. Eckert, "Cyberbullying Detection in Social
Networks Using Deep Learning Based Models; A Reproducibility Study," arXiv. 2018.
[9] S. Agrawal and A. Awekar, "Deep learning for detecting
cyberbullying across multiple social media platforms," arXiv.
2018.
[10] Y. N. Silva, C. Rich, and D. Hall, "BullyBlocker: Towards the
identification of cyberbullying in social networking sites," 2016,
doi: 10.1109/ASONAM.2016.7752420.
[11] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People?
Predictive Features for Hate Speech Detection on Twitter," 2016,
doi: 10.18653/v1/n16-2013.
[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated
hate speech detection and the problem of offensive language,"
2017.
[13] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal
attacks seen at scale," 2017, doi: 10.1145/3038912.3052591.
[14] 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.
[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient
estimation of word representations in vector space," 2013.