Vol-08 Issue 04, April- 2024 Impact Factor: 7.936
ISSN: 2456-9348

iJETRM
International Journal of Engineering Technology Research & Management
www.ijetrm.com

CYBERBULLYING DETECTION AND PREVENTION

Darshini J[1], Daniya Tasneem[2], Chandushree V[3], Chithralekha D N[4], Prof. Velvizhi Ramya R[5]
[*1,2,3,4]UG Student, Computer Science Department, AMC Engineering College, VTU, Bengaluru, India
[5]Assistant Professor, Computer Science Department, AMC Engineering College, VTU, Bengaluru, India

ABSTRACT
This paper presents a novel cyberbullying detection and prevention tool leveraging Long Short-Term Memory (LSTM) networks. Cyberbullying has become a pervasive issue in online platforms, necessitating advanced techniques for timely detection and intervention. Our tool employs LSTM models, renowned for their capability to capture sequential patterns in text data, to analyze and classify potentially harmful content in Social media platforms accurately. By training on large datasets of annotated cyberbullying instances, the LSTM-based system learns to discern nuanced linguistic cues indicative of cyberbullying behavior. Furthermore, the tool incorporates a real-time monitoring mechanism that continuously scans online content, promptly flagging instances of cyberbullying for immediate intervention. Through a combination of natural language processing techniques and deep learning methodologies, our system offers an effective means of combating cyberbullying, fostering safer online environments for users of all ages.

Keywords:
Cyberbullying, Supervised machine learning, Natural language processing, social media.

I. INTRODUCTION

In today's digital age, the prevalence of cyberbullying poses a significant challenge to the safety and well-being of individuals, particularly in online communities and social media platforms. Recognizing the urgent need for effective intervention, we present a pioneering solution: a Cyberbullying Detection and Prevention Tool harnessing the power of Long Short-Term Memory (LSTM) networks. Cyberbullying, characterized by the use of electronic communication to harass, intimidate, or demean others, has emerged as a pervasive issue with far-reaching consequences. Victims often suffer from psychological distress, social isolation, and even physical harm, while perpetrators evade traditional monitoring methods due to the anonymity afforded by the internet. Addressing this complex problem demands innovative approaches that leverage advanced technologies to proactively identify and mitigate instances of cyberbullying.

This paper proposes a tool that capitalizes on the capabilities of LSTM, a type of recurrent neural network renowned for its ability to model sequential data and capture long-term dependencies. By analyzing the textual content of online interactions, LSTM can discern patterns indicative of cyberbullying behavior, distinguishing between harmless exchanges and potentially harmful communications. This deep learning framework empowers our tool to adapt and learn from vast datasets, continually refining its detection capabilities to keep pace with evolving cyberbullying tactics.

The architecture of our Cyberbullying Detection and Prevention Tool comprises multiple layers of LSTM cells, each tasked with processing input text and extracting meaningful features indicative of cyberbullying. Through a process of training on labeled datasets containing examples of cyberbullying and non-bullying interactions, our model learns to recognize subtle nuances in language and context, enhancing its sensitivity and specificity in detecting problematic content. Furthermore, the tool integrates mechanisms for real-time monitoring, enabling swift intervention and preventive measures to safeguard users from harm.

In addition to its robust detection capabilities, our tool prioritizes prevention through proactive interventions aimed at disrupting the cycle of cyberbullying. Leveraging insights gleaned from the analysis of past incidents, the tool offers personalized recommendations and guidance to users, empowering them to navigate online interactions safely and

iJETRM
International Journal of Engineering Technology Research & Management
www.ijetrm.com

responsibly. Through a combination of automated alerts, educational resources, and community support features, our tool fosters a culture of digital civility and mutual respect, cultivating healthier online environments for all.

In summary, our Cyberbullying Detection and Prevention Tool represents a pioneering effort to combat the scourge of cyberbullying using cutting-edge LSTM technology. By harnessing the power of deep learning and real-time monitoring, we aspire to create safer and more inclusive online spaces where individuals can express themselves freely without fear of harassment or intimidation.

II. LITERATURE SURVEY

Shovan Bhowmik and team [1] proposed a model, where they have applied hard voting to create the ensemble frameworks. Because of the hard voting model, if majority of the classifiers in a classifier group predict a tweet as 'offensive', the ensemble model will assess that tweet as 'offensive'. Thus, the prediction results provided by the proposed scheme can validate the result with more confidence than individually applied algorithms. To predict whether the text is 'offensive' or 'non-offensive', after evoking 'BoW' and 'TF-IDF' from the cleaned data, the features have been deployed to the proposed SLE and DLE, and various performance measurement metrics have been measured to evaluate the model performance.

They have built level-based ensemble models to detect cyberbullying from Twitter data. 'BoW', 'TF-IDF' have been generated from noise-free data and fitted to the proposed model. Among various ML algorithms, four well-known classifiers (MNB, LR, DT, LinearSVC) and three ensemble methods (GBoost, AdB and Bagging) have been taken to build their model. To estimate the performance of the framework, they have measured various performance evaluation metrics (Accuracy, F1-Score and AUC). We have achieved 94% accuracy for SLE when TF-IDF ('Word' and 'Unigram') is used to extract features which has surpassed the performance of other models. 70% accuracy has been attained when TF-IDF ('Bi-gram') is considered to retrieve features from tweets. They have also obtained salient performance for other features as well for our SLE model. Moreover, F1-Score and AUC results have been also satisfactory. They have achieved the lowest accuracy for TF-IDF ('Bigram') since N-gram models generally try to find the context of sentences based on bytes, words, syllables, or characters and there are less contextual texts in dataset as it has been collected from Twitter. Moreover, they have tuned to 5000 as maximum features when taking N-gram as a feature value. They have also increased another level and divided the algorithms into two classifier groups for constructing the DLE model. The division has been made in such a way that all well-known ML classifiers are placed in one group and the ensemble methods are placed in another group. Though SLE has achieved higher performance than DLE, for TF-IDF('Bigram'), DLE has achieved 75% accuracy which is greater than SLE. The dataset was partially balanced. To prevent model from bias, they have also validated model by applying different cross-validation techniques (K-Fold, Stratified KFold, Shuffle Split, Stratified Shuffle Split). 10-Fold cross-validation has been applied in this work. Stratified K-Fold and Stratified Shuffle Split have been considered to randomize the dataset into 10 folds so that class wise split can be maintained. Finally, cyberbullying takes many forms such as harassment, flaming, denigration, impersonation, racism, sexism etc. So far, they have only categorized data into two groups. Therefore, it would be great to extend and examine if the proposed model can work for the multi-class classification problem too.

Amirita Dewani and team [2] proposed an advanced preprocessing techniques & deep learning architecture for detection of cyberbullying for Roman Urdu data. The existing research was directed towards mature languages and highlights a huge gap in newly embraced resource poor languages. One such language that had been recently adopted worldwide and more specifically by south Asian countries for communication on social media is Roman Urdu i.e Urdu language written using Roman scripting. To address this research gap, they have performed extensive preprocessing on Roman Urdu microtext. This typically involves formation of Roman Urdu slang- phrase dictionary and mapping slangs after tokenization. They have also eliminated cyberbullying domain specific stop words for dimensionality reduction of corpus. The unstructured data were further processed to handle encoded text formats and metadata/non-linguistic features.

iJETRM
International Journal of Engineering Technology Research & Management
www.ijetrm.com

Furthermore, they performed extensive experiments by implementing RNN-LSTM, RNN-BiLSTM and CNN models varying epochs executions, model layers and tuning hyperparameters to analyze and uncover cyberbullying textual patterns in Roman Urdu. The efficiency and performance of models were evaluated using different metrics to present the comparative analysis. Results highlight that RNN-LSTM and RNN-BiLSTM performed best and achieved validation accuracy of 85.5 and 85% whereas F1 score was 0.7 and 0.67 respectively over aggression class.

Niraj Nirmal and team [3] proposed a method to detect cyberbullying activities on social media. The detection method can identify the presence of cyberbullying terms and classify cyberbullying activities in social network such as Flaming, Harassment, Racism and Terrorism using natural language processing and machine learning algorithms.

So, in this project they focused on making a model on automatic cyberbullying detection in social media text by modelling posts written by bullies on social network. They developed the project using python and web technology. Within that first they search and find the dataset and download it for train the model. After downloading, they will first pre-process the data and then transferred to TFIDF. Then with the help of Naïve bayes, SVM (Support vector machine) and DNN algorithm they train the dataset and generate model separately. Then they developed a web- based application using FLASK framework. They will fetch the real time tweets from twitter and then they apply generated model to these fetched tweets and check the text or images are cyberbullying or not. For all these purposes they are using python as backend, MySQL is database and for frontend HTML, CSS, JavaScript etc.

Approaches of model used are:

1. Naïve Base Model
2. SVM Model
3. DNN Model

The Naive Bayes family of classifiers are simple conditional probabilistic classifiers that work by applying Bayes theorem with naive independence assumptions between the different features.

SVM (Support Vector machine) is a supervised learning algorithm, and is one of the most efficient and universal classification algorithms. Its goal is to find the optimal separating hyperplane which maximizes the margin of training data. Initially the classifier is trained with labelled data before being used to classify the data to test accuracy. Before the data can be used to train our classifier, it is imperative to process it.

Deep Layered Network Architecture Deep neural networks compose computations performed by many layers.

The goal of this project is to the automatic detection of cyberbullying-related posts on social media. Given the information overload on the web, manual monitoring for cyberbullying has become unfeasible. Automatic detection of signals of cyberbullying would enhance moderation and allow to respond quickly when necessary. The main aim of this project is that it presents a system to automatically detect signals of cyberbullying on social media, including different types of cyberbullying, covering posts from bullies, victim and bystanders.

Dorothy L. Espelage and Jun Sung Hong [4] published an article on Prevention and intervention materials, from websites and tip sheets to classroom curriculum, have been developed to help youth, parents, and teachers address cyberbullying. While youth and parents are willing to disclose their experiences with bullying to their health care providers, these disclosures need to be taken seriously and handled in a caring manner. Health care providers need to include questions about bullying on intake forms to encourage these disclosures. The aim of this article is to examine the current-status of cyberbullying prevention and intervention. Research support for several school-based intervention programs is summarized.

III. METHODOLOGY

As shown in Figure 1, with the help of LSTM algorithm, we train the dataset and generate a model which is applied to the fetched real time data from social media platforms.

In the realm of cyberbullying detection, Long Short-Term Memory (LSTM) networks play a crucial role. Let's delve into how they work and their application in identifying cyberbullying behaviour:

iJETRM
International Journal of Engineering Technology Research & Management
www.ijetrm.com

1. Understanding LSTMs: LSTMs are a type of recurrent neural network (RNN) architecture designed to handle sequential data, such as text. Unlike traditional RNNs, LSTMs can capture long-term dependencies and maintain information over extended sequences. They consist of memory cells that allow them to store and retrieve information over time.
2. Temporal Aspects and Sequential Dependencies: In the context of cyberbullying detection, LSTMs excel at understanding the temporal aspects and sequential patterns present in textual content. When analysing social media posts or messages, it's essential to consider the order of words and phrases. LSTMs can model this effectively.
3. Textual Patterns and Behaviour Detection: LSTMs learn to recognize patterns in the input data by processing it sequentially. For cyberbullying detection, they can identify specific linguistic cues, offensive language, and aggressive behaviour. By analysing the context and relationships between words, LSTMs can flag potentially harmful content.
4. Preprocessing and Feature Extraction: Before feeding text data into an LSTM model, preprocessing steps are crucial:
   - Tokenization: Breaking down sentences into individual words or sub-word units.
   - Stop word removal: Eliminating common words (e.g., "the," "and") that don't carry significant meaning.
   - Encoding: Representing words as numerical vectors (word embeddings).
5. Model Architecture: An LSTM model typically consists of:
   - An embedding layer: Converts word indices into dense vectors.
   - LSTM layers: Process the sequential data and capture dependencies.
   - Additional layers (e.g., fully connected layers) for classification.

   The LSTM layers learn to recognize patterns related to cyberbullying.
6. Training and Evaluation: The model is trained on labelled data (cyberbullying vs. non-cyberbullying). During training, it adjusts its weights to minimize the prediction error. Evaluation metrics (such as accuracy, precision, recall, and F1 score) assess its performance.
7. Word Embeddings: Word embeddings (e.g., Word2Vec, GloVe) help LSTMs understand semantic relationships between words. These embeddings capture contextual information, making them valuable for detecting toxic language.

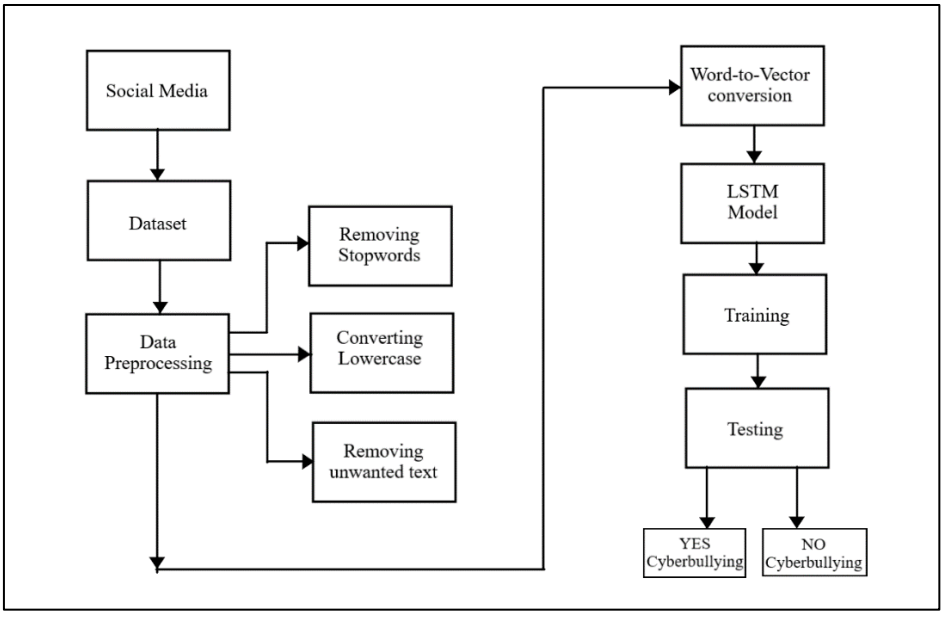

Figure 1 : Architecture of Proposed tool

iJETRM
International Journal of Engineering Technology Research & Management
www.ijetrm.com

IV. CONCLUSION

In conclusion, our utilization of LSTM technology within the realm of cyberbullying detection and prevention represents a significant stride towards fostering a safer and more compassionate online environment. By harnessing the power of deep learning and real-time monitoring, we have developed a robust tool capable of identifying and mitigating instances of cyberbullying with unprecedented accuracy and efficiency. Through proactive intervention and personalized guidance, we aim to empower individuals to navigate digital spaces with confidence and resilience, ultimately fostering a culture of respect, empathy, and inclusion. As we continue to refine and expand our tool's capabilities, let us remain steadfast in our commitment to combatting cyberbullying and creating a world where everyone can thrive free from the fear of harassment or intimidation.

V. REFERENCES

[1] Shovan Bhowmik, Kazi Saeed Alam, and Priyo Ranjan Kundu Prosun, "*Cyberbullying Detection: An Ensemble Based Machine Learning Approach*", Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021). IEEE Xplore Part Number: CFP21ONG-ART; 978-0-7381-11834.

[2] Amirita Dewani, Mohsin Ali Memon and Sania Bhatti, "*Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data*", Dewani et al. Journal of Big Data (2021) 8:160 https://doi.org/10.1186/s40537-021-005507.

[3] Niraj Nirmal, Pranil Sable, Prathamesh Patil, Prof. Satish Kuchiwale, "*Automated Detection of Cyberbullying Using Machine Learning*", International Research Journal of Engineering and Technology (IRJET), Volume : 07, Issue : 12 | Dec 2020.

[4] Dorothy L. Espelage, PhD and Jun Sung Hong, PhD, "*Cyberbullying Prevention and Intervention Efforts: Current Knowledge and Future Directions*", The Canadian Journal of Psychiatry / La Revue Canadienne de Psychiatrie 2017, Vol. 62(6) 374-380.