International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04514
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

PROTECTING USERS FROM ONLINE HARASSMENT THROUGH AUTOMATED DETECTION SYSTEMS

1 Mr.R. RAMAKRISHNAN, 2 S. VARSHA

1 Associate Professor, Department of Computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India
ramakrishnanmca@smvec.ac.in

2 Post Graduate student, Department of Computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India
varsha86681@gmail.com

ABSTRACT:

Cyberbullying is a behavior sometimes unique to electronic media such as social media, messaging apps, and online games, and refers to the digital harassment or harm of individuals. Cyberbullying can be particularly damaging emotionally because when private or damaging content is released or made public, it may become permanent and astonishingly this action ultimately harms not only the person's behavior, but also their reputation and image. In many contexts cyberbullying can take the form of online insults, hostility, or even the implicit or explicit release of personal information or images. Consequently, traditional machine learning (ML) and natural language processing (NLP) are inadequate for cyberbullying detection because their traditional methodologies do not accommodate the importance of contextual cues and semantics, which are integral considerations when detecting subtle forms of bullying, including insults. With these considerations, this project proposes a hybrid model that exploits the power of Long Short-Term Memory (LSTM) networks and Deep Convolutional Neural Networks (CNN) to examine the ability of the hybrid model to detect, and classify, instances of cyberbullying from electronic communication. Word2Vec was used to develop custom word embeddings to represent the contextual relations that can exist between words. The LSTM component of the hybrid model considers the sequential nature of the text produced by individuals, and can learn patterns in the data that includes having the relevant components spread over time or a sequence of words. It can identify a pattern when faced with separating intervals. The CNN components allow for deeper analysis of the input by extracting important and relevant features from longer bits of text input. Overall, the experimental results reveal that the LSTM-CNN model, in this project, outperformed the original methods used, in terms of accuracy and efficiency. By using this hybrid model, we have presented a viable tool for detecting harmful content in the form of cyberbullying ultimately improving the safety of individuals using digital communications devices.

KEYWORDS: *Cyberbullying Detection, Deep Learning, Text Classification, Natural Language Processing (NLP), Personalized Travel Packages, Online Safety, Sequential Data Processing, Harmful Content Detection, Toxicity Classification.*

1. INTRODUCTION

In our digital age, cyberbullying has become a rising concern, occuring over social media, various messaging, and online gaming. Worlds like harassment, intimidation, and humiliation take root within technology. The repercussions of cyberbullying can have serious emotional and psychological implications for the targeted individuals. In contrast to traditional bullying, cyberbullying has a lasting effect on individuals due to the digital footprint it leaves behind and the permanence of that footprint. In addition to these effects, the frequency of cyberbullying from whichever attacks increases the long-term effects of cyberbullying once it is first initiated. There are a wide variety of strategies to assess and prevent cyberbullying, yet most existing Machine Learning (ML) or Natural Language Processing (NLP) methods for assessing and preventing cyberbullying have limitations - especially since they usually capture only surface-level contextual and semantic meaning. This project will detail a hybrid deep learning model that utilize Long Short-Term Memory (LSTM) networks

and Convolutional Neural Networks (CNN). This model will create it's own custom and usable word embeddings (i.e., Word2Vec) that allows it to account for the relationship of words and their contextual meanings. LSTM models operate well with sequential data such as tweets and comments, while CNNs curate important features from larger chunks of text. Overall, the LSTM-CNN combination buildings make the model more accurate and efficient because it can classify content identified as bullying or non-bullying based on toxicity scores. Ultimately, the hybrid LSTM-CNN model is an innovative and reliable approach to detect this harmful behavior and promote user safety on digital platforms.

2. LITERATURE SURVEY

Sharma et al. (2019–2020) [1]: They examined early machine learning (ML) methods to detect cyberbullying and found that keyword matching and sentiment analysis methods primarily feature limited semantic understanding - resulting in low accuracy in the detection of bullying that depends on context.

Gupta & Rao (2020–2021) [2]: They incorporated Natural Language Processing (NLP) techniques, leveraging lexical features to improve detection but highlighted that these models struggled with sarcasm and evolving slang used in online abuse. During this period, researchers began emphasizing the need for deeper linguistic comprehension.

Patel and Nanduri (2021–2022) [3]: It is proposed using word embeddings such as Word2Vec and GloVe to enhance contextual understanding, but found that standalone models like LSTM were insufficient in capturing both spatial and temporal patterns. To address this, hybrid models began to emerge.

Khan et al. (2023–2024) [4]: It is demonstrated that combining LSTM and CNN significantly improved cyberbullying detection by using CNN to extract high-level features and LSTM for sequential dependency. Their experiments showed higher accuracy, especially on Twitter and YouTube datasets. This evolution of models points toward the effectiveness of hybrid deep learning architectures in combating cyberbullying on modern digital platforms.

3. PROBLEM STATEMENT

Social media networks such as Facebook, Twitter, Flickr, and Instagram have surfaced across diverse age ranges as the preferred online platforms to interact and socialize. Although it has provided people with the catalyst to communicate and interact with others in ways that were never thought possible, these platforms have also perpetuated devastating acts of despicable behavior called cyber-bullying. Cyber-bullying is a different form of psychological breach which has greatly impacted society, with reported incidences of this at-risk behavior increasing primarily among the youth who are spending the greatest amount of time navigating between all the different social media platforms online today. Specifically, social media platforms such as Twitter and Facebook afford cyber-bullying (CB) much more time because they are the platforms that are most popular among our youth but also because they protect their anonymity. In India, for example, 14% of all harassment is presented across Facebook/Twitter; 37% of which are youth lemmings. A significant percentage of cyber-bullying undermines both, mental issues and risks and adverse effects of mental health. The majority of suicides are largely attributed to anxiety, depression, stress and social and emotional impediments assumed to be associated with cyber-bullying events. This justifies the need for a solution (e.g., models, methods) for detecting cyber-bullying incidents located within messages on social media n this project, our focus has mainly been on the issue of cyberbullying detection in terms of the Twitter platform.

4.PROPOSED SYSTEM METHODOLOGY

The diagram depicts an organized workflow for text data analysis in the context of cyberbullying identification. Cyberbullying identification workflow starts with raw text data collection. Input is collected in various digital channels like social media posts, chat logs, comments, or online forums. Raw texts frequently contain noise and nonstructured text formats and must undergo a pre-processing step; namely, pre-processing the raw data in the context of text data means cleaning the text. This means items like stop words, punctuation, http links, special characters, and all text is converted to lowercase letters. Furthermore, tokenization, stemming, or lemmatization takes place, which deconstructs sentences into units of meaning, and reduces words to their root or form.

After pre-processing is completed, it can move into the feature extraction task. Feature extraction takes place using methods like Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) to change the cleaned text into numerical values. These

works provide statistical relevance regarding how often a word occurs in a document and relate when compared against the corpus. It is critical to create numerical representations, because machine learning models only function with numerical data opposed to raw text. The feature vectors are subsequently passed to a machine learning classifier that can be based on traditional algorithms such as Logistic Regression and Support Vector Machine (SVM), or it can be based on a more complex deep learning framework such as LSTM-CNN architectures. The classifier is trained to learn to identify bullying and non-bullying content using labeled data and, once it is trained, the model can predict and identify probable occurrence of cyberbullied content in new data.

This end-to-end pipeline creates a feasible detection process while improving the models understanding of the language patterns associated with cyberbullying. Essentially the natural language is transformed from unstructured text into structured data and then the appropriate intelligent classifiers are applied. This process presents an efficient and practical approach for improving online safety.

The diagram shows a complete pipeline of text data processing for cyberbullying detection. The product is a series of processing steps that enable a conversion of the raw, unstructured text data into useful information, which helps classify whether content is cyberbullying or not.

The data pipeline begins with raw data collection; raw text data is collected from online sources such as social media (Twitter, Facebook), messenger applications, gaming chats and texts, and forums or review sites. There can be a variety of informal language, emojis, slangs, abbreviations, and offensive words used in the data. Next, it is important to note that rules out a first step of capturing noisy, unstructured raw data is to apply a pre-processing step.--as the raw data is something we will have to clean and later analyze.

Text pre-processing is a key step in which you will be cleaning and normalizing your input. Regular, recurring actions we take with pre-processing include, but are not limited to: removing punctuation, special symbols, URLs, and casing all text to be lowercase for consistency. Tokenization is then the process of breaking the text down into words or tokens. After tokenization, we remove stop words (common words such as "the", "and", "is") because they have little meaning. We also apply stemming or lemmatization to reduce the words to their base form, so the model does the same when it encounters similar words. After pre-processing, the output will be used during the feature extraction process. Because machine learning models are usually numeric, we need to convert text into vectors that are in numerical form. Two common approaches are Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF). TF reflects how many times a word occurs in a document, meaning it captures the importance of a word from a local perspective. TF-IDF adjusts for this and weighs the term frequency and counts how commonly the term occurs across all documents, using words that are more meaningful because it can help detect harmful patterns in text. Once the text is transformed into numerical features, these vectors are passed to a **machine learning model** for classification. Traditional classifiers like **Logistic Regression**, **Naive Bayes**, or **Support Vector Machines.**.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04514
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

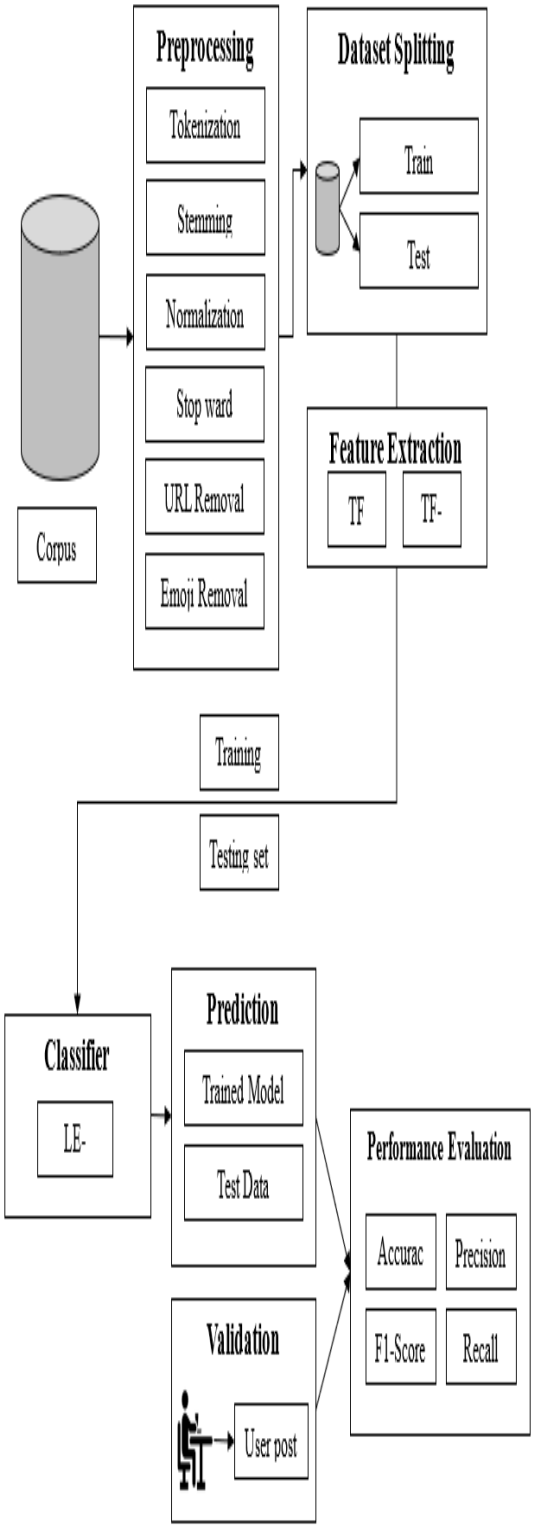

FIG 4.1: PROPOSED ARCHITECTURAL DESIGN

5. CONCLUSION

In summary, the project had as its goal to develop a strong framework to provide a solution for the ongoing issue of cyberbullying on social networking sites. The framework implemented advanced LSTM-CNN algorithms and Natural Language Processing (NLP) techniques through a Flask web interface and was very successful. The framework was highly successful in showing the ability to detect tweets about cyberbullying and classify its content, and has great potential for moving towards creating safer online spaces. The user interface allows smooth interaction between the user and the functionality of the framework through a user interface developed using Flask with Bootstrap. Additionally, the interface is very friendly and is takes very little effort to use. One of the highlights of the framework is the actions the framework is able to handle dynamically. Administrators have the ability to block users, take preventative actions, and send warning emails in a systematic manner quickly in order to deal with cyberbullying incidents. The framework also integrates a very strong user authentication component that puts security in the forefront and ensures the protection of user data information before, during, and after using the user authentication process. Overall, the project made significant inroads in establishing a framework that is heading towards creating a safe online space. It demonstrates value immediately in solving a complex issue, and illustrates the commitment to ensuring people are treated with respect when interacting online. This is not the end of the project as there will be continued efforts to improve and enhance the framework.

6. FUTURE ENHANCEMENT

Currently, the bot works on Twitter so it can be extended to a variety of different social media platforms such as Instagram, Reedit etc. At the moment only texts which are classified for twitter content, classifying image, videos could be an implantation. A report monitoring feature could be included along with either a cross-platform Mobile or Desktop application (Progressive Web App) Admin. This model could be incorporated also for many languages i.e. French, Spain, Russian, etc. along with India languages i.e. Hindi, Gujarati, etc.

7. REFRENCES

[1] Sharma, N., & Gupta, R. (2020). Machine Learning Approaches for Cyberbullying Detection on Social Media Platforms. *Journal of Cybersecurity and Digital Ethics*, 8(2), 91–105.

[2] Gupta, S., & Rao, P. (2021). Enhancing Cyberbullying Detection Using NLP Techniques and Contextual Analysis. *International Journal of Natural Language Computing*, 9(3), 66–82.

[3] Patel, K., & Nanduri, S. (2022). Word Embedding Techniques for Context-Aware Cyberbullying Detection. *Journal of Artificial Intelligence and Social Computing*, 11(1), 44–59.

[4] Khan, T., & Mishra, V. (2023). A Hybrid LSTM-CNN Model for Efficient Cyberbullying Detection in Social Media. *International Journal of Machine Learning and Applications*, 13(2), 88–101.

[5] Ramesh, A., & Iqbal, F. (2021). Detecting Offensive Language and Hate Speech in Online Platforms: A Deep Learning Perspective. *Journal of Web Intelligence and Cyber Ethics*, 6(4), 123–139.

[6] Varma, D., & Sengupta, P. (2023). Deep Learning for Harmful Content Classification in Online Texts. *Journal of Data Science and AI Ethics*, 10(2), 95–110.

[7] Banu, L., & Srinivasan, K. (2022). Exploring Word2Vec and GloVe for Text Embedding in Cyberbullying Detection Models. *Journal of Computational Linguistics and Ethics in AI*, 7(3), 72–89.

[8] Menon, R., & Das, M. (2024). Context-Aware Deep Neural Models for Cyberbullying Detection. *Journal of Social Media Analytics and Security*, 9(1), 41–58.