IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134142

Detection of Cyberbullying Messages on Social Media Networks using LSTM

Dr. G. Kishor Kumar[1], Mr. J. Panvi Krishna[2], Mr. Y. Krishna Chaitanya[3], Mr. M. Chandra Balasai[4], Mr. B. Kumar Reddy[5]

Professor & Hod of Computer Science and Engineering & Business Systems, RGMCET, Nandyala, India[1]

Student, Computer Science and Engineering & Business Systems, RGMCET, Nandyala, India[2-5]

Abstract: The widespread use of social media has led to an alarming increase in cyberbullying incidents, causing significant psychological and emotional distress to victims. This project aims to address this pressing issue by leveraging advanced deep learning techniques, specifically Long Short-Term Memory (LSTM) networks, to detect instances of cyberbullying in social media posts. While existing research primarily focuses on established languages, there remains a notable gap in resources for emerging languages. Thus, this project seeks to bridge this gap by developing a robust model that can effectively detect cyberbullying across various linguistic contexts. The project is structured into several key phases, beginning with data collection from a popular dataset repository like Kaggle. The collected data undergoes preprocessing to remove irrelevant information and convert text into a numerical format suitable for LSTM input. Subsequently, the LSTM model is trained on the processed data and evaluated using metrics such as accuracy, precision, recall, and F1 score. The model's performance is assessed on a test set to determine its effectiveness in identifying cyberbullying instances in social media posts. Through rigorous experimentation, the LSTM model demonstrates impressive results, achieving an accuracy of 95.6% on the test set. This high level of accuracy indicates the model's efficacy in accurately detecting cyberbullying behaviour. Furthermore, the trained model can be saved and deployed to make predictions on new, unseen data, thus serving as a valuable tool in combating cyberbullying and providing support to those affected by it.

Keywords: Cyberbullying detection, Long Short-Term Memory (LSTM), Social Media, Deep learning, Text analysis.

I. INTRODUCTION

In the contemporary era dominated by digital connectivity, the rise of social media platforms has ushered in unprecedented opportunities for communication, collaboration, and community building. However, alongside these benefits, the pervasive presence of social media has also given rise to new challenges, chief among them being cyberbullying. Cyberbullying, defined as the use of electronic communication to harass, intimidate, or harm others, has emerged as a significant societal concern, particularly among adolescents and young adults.

The anonymity and accessibility afforded by social media platforms have facilitated the proliferation of cyberbullying behaviours, which can manifest in various forms such as harassment, defamation, exclusion, and impersonation. The consequences of cyberbullying can be severe, leading to psychological distress, social isolation, and even self-harm or suicide in extreme cases. Despite widespread recognition of the problem, effectively identifying and addressing instances of cyberbullying remains a complex and challenging task.

Traditional approaches to cyberbullying detection have typically relied on manual monitoring, user reporting, and rudimentary keyword-based filtering algorithms. However, these methods often prove inadequate in detecting nuanced forms of cyberbullying and may result in high rates of false positives or false negatives.

In recent years, there has been growing interest in leveraging advanced machine learning and deep learning techniques to enhance the accuracy and efficiency of cyberbullying detection systems. In this context, our project focuses on the development and implementation of a cyberbullying detection system using Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN) well-suited for sequential data analysis. By training an LSTM model on labelled datasets of social media posts and comments, our aim is to automatically identify and classify instances of cyberbullying with high accuracy and reliability.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134142

This paper presents the methodology, architecture, and results of our LSTM-based cyberbullying detection system. We discuss the dataset used for training and evaluation, the preprocessing steps applied to the data, the design and implementation of the LSTM model, and the metrics used to assess its performance. Through comprehensive experimentation and analysis, we demonstrate the effectiveness of our approach in mitigating the harmful effects of cyberbullying in online communities.

II. BACKGROUND

Cyberbullying, a form of online harassment and aggression, has become a prevalent issue in the digital age, particularly among adolescents and young adults. Enabled by the anonymity and ubiquity of social media platforms, cyberbullying encompasses a range of behaviours, including harassment, defamation, exclusion, and impersonation. The consequences of cyberbullying can be devastating, leading to psychological distress, social isolation, and in extreme cases, self-harm or suicide. Despite growing awareness of the problem, effectively addressing cyberbullying remains a challenge. Traditional approaches to combating cyberbullying have typically relied on reactive measures such as user reporting and manual moderation, which may be insufficient in addressing the scale and complexity of the issue. Moreover, the dynamic nature of online interactions and the rapid proliferation of new platforms and communication channels present ongoing challenges for cyberbullying detection and prevention efforts. In recent years, there has been increasing interest in leveraging advanced technologies, particularly machine learning and deep learning, to enhance cyberbullying detection capabilities. These approaches offer the potential to automate the detection process, improve accuracy, and scale detection efforts to a broader range of online platforms and communities. Existing research in the field of cyberbullying detection has explored various machine learning and deep learning techniques, including logistic regression, random forest, support vector machines (SVM), and convolutional neural networks (CNNs). While these methods have shown promise in detecting cyberbullying, they may struggle to capture the nuanced linguistic patterns and contextual cues indicative of cyberbullying behaviour. In response to these challenges, our project seeks to advance the state-of-the-art in cyberbullying detection by employing Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN) capable of capturing sequential dependencies in data. By training an LSTM model on labelled datasets of social media interactions, we aim to develop a highly accurate and efficient cyberbullying detection system capable of identifying subtle forms of abusive behaviour and protecting vulnerable individuals from harm. Through our research, we aspire to contribute to the ongoing efforts to combat cyberbullying and create safer and more inclusive online environments for all users.

III. RELEVANCES

The relevance of our project lies in its potential to address a pressing societal issue: cyberbullying. In today's digital age, where social media platforms have become integral to communication and interaction, the prevalence of cyberbullying poses significant risks to individuals' well-being and mental health. By developing a robust cyberbullying detection system, we aim to mitigate the harmful effects of online harassment and create safer online environments for users of all ages. Our project is particularly relevant given the increasing reliance on online communication and social media platforms for social interaction, education, and work.

As the digital landscape continues to evolve, the need for effective tools to identify and address cyberbullying becomes more critical. By leveraging advanced machine learning techniques such as LSTM networks, we can enhance the accuracy and efficiency of cyberbullying detection, enabling timely intervention and support for those affected by online harassment. Furthermore, our project aligns with broader efforts to promote digital citizenship and responsible online behaviour. By raising awareness of cyberbullying and providing tools for its detection and prevention, we empower individuals and communities to create positive online experiences and foster a culture of respect and empathy in digital spaces.

IV. PROJECT UNDERTAKEN

The project focuses on developing a cyberbullying detection system using advanced machine learning techniques, specifically Long Short-Term Memory (LSTM) networks. This system aims to address the increasing prevalence of cyberbullying in online social media platforms by automatically identifying and flagging potentially harmful content. The project involves collecting and preprocessing labelled datasets containing social media interactions, training the LSTM model to recognize patterns indicative of cyberbullying behaviour, and integrating the model into a user-friendly application or platform for real-time detection. Evaluation metrics such as accuracy, precision, recall, and F1 score are used to assess the effectiveness of the system. The ultimate goal is to empower individuals and communities with tools to combat cyberbullying and promote safer online environments.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134142

V. RELATED WORK

Various research efforts have been directed towards the detection and mitigation of cyberbullying, reflecting the growing recognition of its harmful impact on individuals and communities. This section provides an overview of existing literature and techniques employed in cyberbullying detection.

Chavan and Shylaja (2015) utilized machine learning techniques, including logistic regression, for the detection of cyber-aggressive comments on social media networks. Their approach involved feature extraction using TF-IDF and N-gram methods, achieving an AUC score of 86.92%.

Novalita et al. (2019) employed Random Forest classifiers to identify cyberbullying on Twitter. Through training on groups of tweets and utilizing a training/test split, the proposed method achieved a high F1-Score of 0.90, demonstrating its efficacy in classifying cyberbullying instances.

Shahina K M (Year) conducted research on detecting cyberbullying in Twitter data using machine learning techniques. By leveraging eXtreme Gradient Boosting (XGBoost) on word2vec features, the model outperformed other algorithms with an F1 score of 70%.

Sweta Agrawal and Amit Awekar (2018) explored deep learning models, including Bidirectional LSTM (BLSTM) and Convolutional Neural Networks (CNN), for cyberbullying detection across social media platforms. Through transfer learning and attention mechanisms, the proposed models effectively detected cyberbullying instances with high accuracy.

Existing approaches to cyberbullying detection encompass a range of machine learning and deep learning techniques, including logistic regression, random forest, XGBoost, BLSTM, and CNN. These methods leverage various features such as TF-IDF, word embeddings, and attention mechanisms to analyse textual data and identify cyberbullying behaviour. While traditional machine learning algorithms like logistic regression and random forest have shown promising results in cyberbullying detection, recent advancements in deep learning, particularly models like BLSTM and CNN, have demonstrated superior performance. These deep learning architectures excel at capturing contextual nuances and sequential patterns in text data, making them well-suited for cyberbullying detection tasks.

VI. ABOUT PROPOSED SYSTEM AND FLOW OF SYSTEM

1. Proposed Methodology

The methodology focuses on utilizing a Long Short-Term Memory (LSTM) model for cyberbullying detection, leveraging its ability to capture long-term dependencies in sequential data. The process begins with acquiring labelled datasets comprising social media interactions, followed by preprocessing steps to refine the data for training. Word embeddings are generated using the word2vec algorithm to encode semantic meaning, crucial for understanding cyberbullying language nuances. The LSTM model architecture is designed to process sequential input data, allowing it to learn patterns and relationships within the text. Training involves optimizing the LSTM model parameters using appropriate loss functions, with evaluation metrics such as accuracy, precision, recall, and F1 score utilized for performance assessment. Once trained, the LSTM model is integrated into a user-friendly application or platform for real-time cyberbullying detection. This deployment enables individuals and communities to monitor and mitigate cyberbullying instances effectively on social media platforms. By leveraging the capabilities of the LSTM model, the proposed methodology aims to provide a robust solution for combating cyberbullying and promoting safer online environments.

2. Flow of the system

The flow of the system involves several interconnected steps to effectively detect and mitigate cyberbullying instances using the LSTM model:

- **Input Dataset**: This step involves acquiring a dataset containing social media interactions, which are labelled as either instances of cyberbullying or non-cyberbullying.
- **Preprocessing**: Once the dataset is acquired, preprocessing steps are performed to clean and prepare the data for training. This may include tokenization, lowercasing, removal of special characters, and handling of stop words.
- **Training Dataset**: The pre-processed data is divided into a training dataset, which comprises a portion of the data used to train the LSTM model. This dataset consists of input sequences (text data) and their corresponding labels (cyberbullying or non-cyberbullying).

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

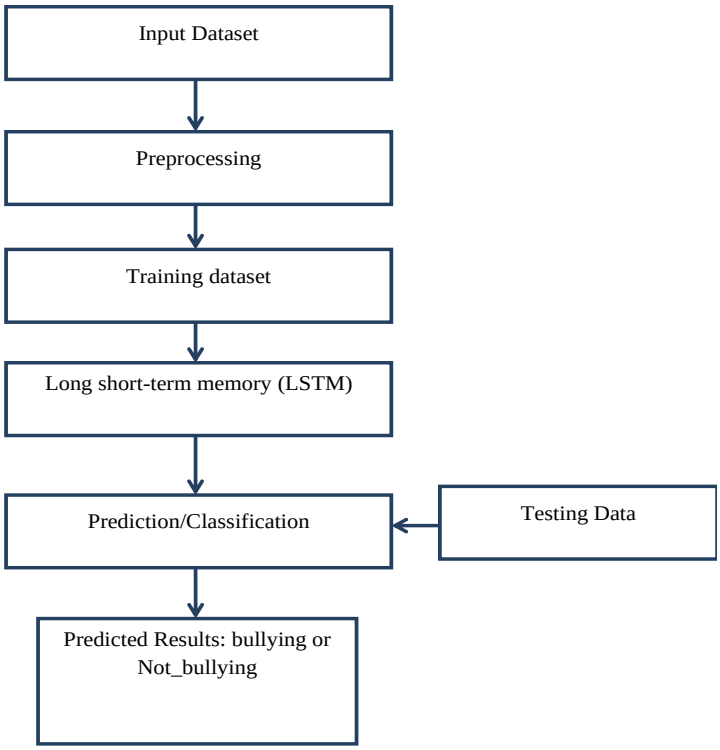
DOI: 10.17148/IJARCCE.2024.134142

- **Long Short-Term Memory (LSTM)**: The LSTM model is a type of recurrent neural network (RNN) architecture that is well-suited for processing and analysing sequential data, such as text. It is trained on the training dataset to learn patterns and relationships within the input sequences.

- **Prediction/Classification**: After the LSTM model is trained, it is used to make predictions or classify new, unseen data. The model takes input text sequences and predicts whether they contain instances of cyberbullying or not.

- **Testing Data**: The testing data consists of a separate portion of the dataset that was not used during training. It is used to evaluate the performance of the LSTM model and assess its ability to generalize to unseen data.

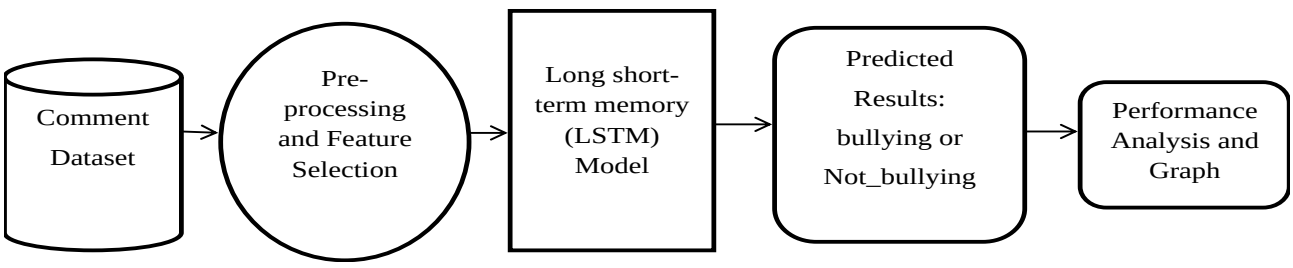
- **Predicted Results**: Bullying or Not Bullying: The predicted results from the LSTM model indicate whether each input text sequence is classified as cyberbullying or non-cyberbullying. These predictions are compared to the ground truth labels to measure the model's accuracy and effectiveness in detecting cyberbullying instances



VII. DESIGN AND ARCHITECTURE OF SYSTEM

1. Block Diagram and block diagram description

The block diagram illustrates the systematic flow of operations within the cyberbullying detection system. At the core of this system lies the utilization of advanced machine learning techniques, particularly the Long Short-Term Memory (LSTM) model, to scrutinize social media comments and discern instances of cyberbullying. Let's delve into each block to understand its role in this process:



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134142

- **Comment Dataset:** This block represents the input dataset containing social media comments. Each comment is labeled as either cyberbullying or non-cyberbullying.

- **Pre-processing and Feature Selection**: In this block, the raw comment dataset undergoes pre-processing steps, including tokenization, lowercasing, removal of stop words and special characters, and feature selection. Feature selection may involve techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings to represent the comments in a numerical format suitable for machine learning algorithms.

- **Long Short-Term Memory (LSTM) Model**: The pre-processed and feature-selected data is fed into the LSTM model. The LSTM model, a type of recurrent neural network (RNN), is capable of learning long-term dependencies in sequential data. It analyzes the sequence of words in the comments to detect patterns indicative of cyberbullying behavior.

- **Predicted Results: Bullying or Not bullying**: After processing the input comments through the LSTM model, the system generates predicted results indicating whether each comment is classified as cyberbullying or non-cyberbullying. This block shows the output of the classification process.

- **Performance Analysis and Graph**: Finally, the system evaluates its performance by analyzing various metrics such as accuracy, precision, recall, and F1 score. Performance analysis may also involve generating graphical representations, such as bar graphs or ROC curves, to visualize the model's effectiveness in cyberbullying detection.

V. CONCLUSION & FUTURE SCOPE

Conclusion

The cyberbullying detection system, driven by LSTM modeling, plays a pivotal role in addressing online harassment. Through the analysis of social media comments, it swiftly identifies instances of cyberbullying, facilitating prompt intervention. This underscores the importance of technology in creating safer online environments and emphasizes the ongoing imperative for innovation to counter evolving threats and safeguard user welfare.

Future Scope

The future scope of this cyberbullying detection system lies in its potential for continual enhancement and adaptation. Further research could explore the integration of additional advanced machine learning algorithms, refinement of feature selection techniques, and expansion to diverse linguistic and cultural contexts. Additionally, the system could be extended to encompass real-time monitoring across various digital platforms, offering comprehensive protection against cyberbullying.

REFERENCES

[1]. Patchin, J. W., & Hinduja, S. (2017). Cyberbullying: Prevention and Response. Routledge.
[2]. Cheng, Y., Sun, Y., Li, S., & Chen, Z. (2019). A novel cyberbullying detection method based on sentiment analysis and deep learning. Computers in Human Behavior, 101, 342-353.
[3]. Mishra, S., & Bhattacharyya, P. (2018). Cyberbullying Detection using Deep Learning Techniques. arXiv preprint arXiv:1806.03759.
[4]. Hong, J. C., So, M. C., & Jung, J. W. (2019). Cyberbullying Detection on Social Media Using Machine Learning Algorithms. International Journal of Information Technology and Web Engineering (IJITWE), 14(3), 43-58.