IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

NOVEL APPROACH FOR CYBERBULLY DETECTION USING NEURAL NETWORKS

[1]Karyada Vaeshnavi, [2]Sree Lasya Balaji, [3]Padala Rishi, [4]Panaganti Koushik, [5]Sreedhar Bhukya

Student, Student, Student, Student, Student, Professor

Department Of Computer Science and Engineering

Sreenidhi Institute of Science and Technology,Hyderabad, India

Abstract - Technology is advancing rapidly today. These advances in technology have significantly changed the way people interact by adding a new dimension to communication. However, even though technology helps us in many aspects of life, it also comes with various consequences that affect people either positively or negatively. Cyberbullying is one such consequence. Cyberbullying is a crime in which a criminal uses hate speech and online harassment to target a person, causing the victim to suffer negative emotional, social, and physical consequences. We suggested a novel deep neural network-based technique for the identification of cyberbullying in order to solve this issue. When compared to current techniques, Convolutional Neural Networks are utilized to provide superior outcomes.

IndexTerms - Cyberbulling,SVM, Long Short-Term Memory, Deep Neural Network, Convolutional Neural Network

I. INTRODUCTION

INTRODUCTION

Cyberbullying involves the continual harassment or torment of an individual through methods such as social media platforms. This conduct aims to intimidate and degrade the victim. Some examples include spreading false information or humiliating

photos of the victim, sending direct abusive messages, or impersonating the victim to sendunwanted messages on their behalf. These are just some instances of cyberbullying that can occur on social media. Online abuse has led to a spike in

cyberbullying, resulting in some student suicides.

Numerous studies have examined detecting attempted cyberbullying on social media based on provided data. However, technological advancements along with the digitalization of relationships have significantly impacted centennials, making it essential to maintain a social media presence. Despite the entertainment social media offers, cyberbullying has been identified as a real problem in Malaysia where these centennials are victims.

This research topic would be valuable since it categorizes cyberbullying intent within tweets. Because cyberbullying can take many different forms, the models developed in this study make use of supervised machine learning[1][4], namely support vector machines and Naive Bayes. Furthermore, a recurrent neural network called Long Short-Term Memory (LSTM) was also implemented. Additionally, the final product will not only be a predictive model but will also provide output to decision-makers in the form of a bag of words.

The present study seeks to examine the issue of cyberbullying, which is becoming increasingly worrisome in this age of technology. The first objective of this project is tackling the challenge of preprocessing data into a format that can be efficiently analysed for data analytics. This involves utilizing appropriate data cleaning procedures, addressing missing data, and ensuring the data is reliable and consistent. Another goal of this research is developing a suitable text classification algorithm that can accurately identify and categorize instances of cyberbullying intent. This requires selecting appropriate features, creating a functional algorithm, and training the model with relevant datasets. When analysing text data, the model should be able to detect patterns and contexts that are indicative of cyberbullying. Ultimately, the model's performance on real-world scenarios will be assessed using appropriate

measures, such as accuracy, precision, recall, F1 score, etc. Additionally, various factors such as data preprocessing, feature selection and model parameters will be examined to study their impact on model's performance.

The objectives of the research are critical in addressing the issue of cyberbullying, as they aim to develop effective methods for identifying and preventing cyberbullying, which is an increasing problem in the digital world. The research findings can be utilized to create a safer and more positive online environment that promotes responsible online behaviour and fosters improved digital interactions.

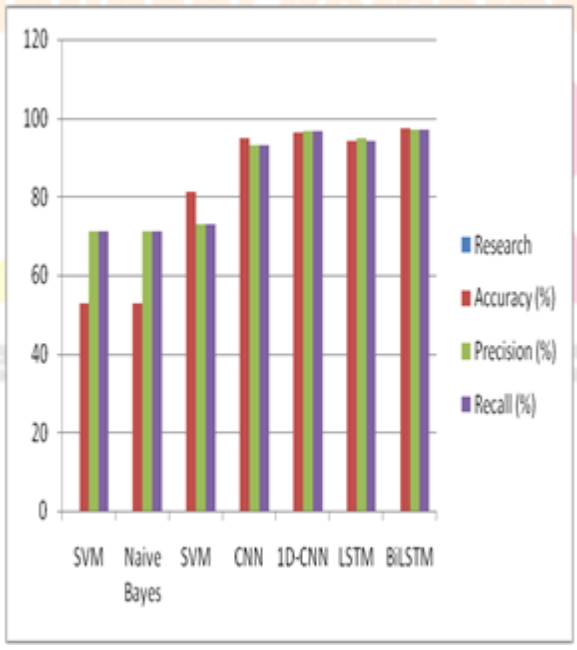
II. LITERATURE SURVEY

A. EXISTING SYSTEM

To understand previous research, many journal publications about detecting cyberbullying using data science techniques were studied. Additionally, earlier models were examined and evaluated as described below. The table below shows some related prior research on identifying cyberbullying using data science methods, where SVM and Naive Bayes - two well-known classical machine learning techniques - were used. Since the SVM and Naive Bayes results in Table-1 were inadequate, this study improved the evaluation outcomes for the machine learning SVM and Naive Bayes algorithms [6]. Furthermore, LSTM is utilized in this study to compare the results of deep learning and conventional machine learning.

table-1: result of prior methods

Models	Research	Accuracy (%)	Precision (%)	Recall (%)
SVM	(Dalvi, Chavan & Halbe, 2020)	52.7	71.0	71.0
Naive Bayes		52.7	71.0	71.0
SVM	(Al-Ajlan & Ykhlef, 2018)	81.3	73.0	73.0
CNN		95.0	93.0	93.0
1D-CNN	(Ghosh, Chaki & Kudeshia, 2021)	96.3	96.5	96.5
LSTM		94.1	94.8	94.3
BiLSTM		97.4	97.0	97.0



graph-1: graphical representation of table-1

B. Disadvantages of the existing system

Traditional machine learning techniques like Support Vector Machines (SVM) or Random Forest are not designed to capture temporal dependencies, which are important when analysing sequences of text. This makes them less suitable for cyberbullying detection, where the context provided by a sequence of messages is key. Additionally, traditional techniques do not handle sequential data well. Cyberbullying often involves a series of connected messages, so considering the context is critical for accurate detection. Another limitation is that traditional methods usually require fixed-length input. Since tweets vary in length, these techniques may struggle to handle such variable-length inputs, again impacting cyberbullying detection accuracy. Finally, conventional machine learning algorithms are sensitive to noisy data. Tweets often contain "noise" like misspellings, grammatical errors, and informal language [7]. This could further degrade the performance of traditional techniques for detecting cyberbullying in tweets.

In summary, key disadvantages of conventional machine learning approaches for cyberbullying detection in tweets are: limited ability to capture temporal context, handle variable-length sequential data, and deal with noisy informal text. More advanced techniques are needed to address these challenges.

C. Deep Neural Network based models:

DNN Deep neural networks such as convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and bidirectional LSTMs (BLSTMs) have been utilized for cyberbullying detection. These models differ in the complexity of their neural architectures. CNNs are commonly used for image classification, text classification [3], and sentiment analysis. LSTMs are useful for learning long-term dependencies due to their internal memory. BLSTMs can encode information in both forward and backward directions, increasing the input data for the network [2]. Most models have similar layers except for the unique neural architecture layer. The embedding layer processes fixed-length word sequences. There are two dropout layers to prevent overfitting - one before and one after the neural architecture. Finally, there is a dense output layer with neurons equal to the number of classes.

CNNs perform well on data with high locality where nearby words are weighted more heavily. Their short length makes them suitable for cyberbullying text. The input text is converted to integer sequences representing unigrams.
Preprocessing includes converting emoticons to words, removing non-Latin characters, and stripping metadata. Hashtags and mentions become binary features. The tokenized and lowercased text is encoded into integers.

BiLSTMs train two LSTMs on the input sequence, providing additional context. This results in faster, fuller learning. The first layer receives the normal sequence, while the second layer receives a reversed copy. Bidirectional LSTMs are useful where the whole input context aids interpretation, like in speech recognition. They may improve results for certain domains, but are not always appropriate.

D. Convolutional Neural Networks (CNNs):

Convolutional Neural Networks (CNNs) have shown good performance on data with high locality, where words are given more importance based on their surrounding context. We are attempting to prioritize short texts that exhibit tendencies of cyber bullying. We utilized CNNs that take as input sequences of integer representations of unigram text. The text handling included changing over emojis into words and eliminating non-Latin characters. We additionally eliminated normal URL parts (for example names of famous sites), metadata in the body text (for example "RT :"), and different stage explicit elements. Hashtags and @-makes reference to were decreased to twofold pointers. The text was lowercased and tokenized utilizing NLTK's TweetTokenizer. The tokenized text was then encoded as whole numbers, protecting the first word request [3][5].

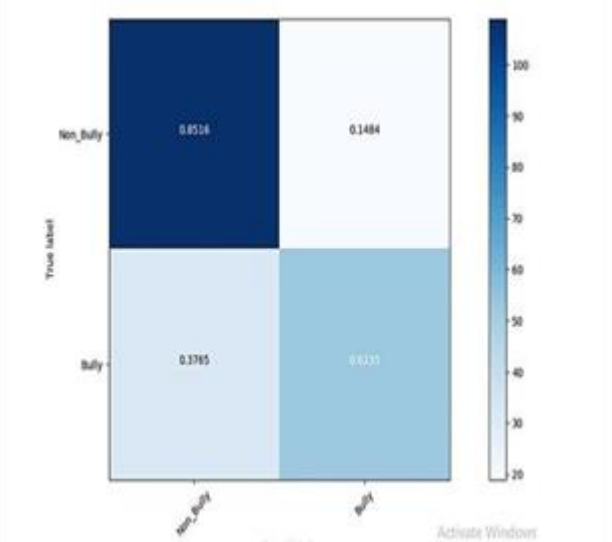


fig 1. the confusion matrix of cnn

The confusion matrix shows the ways in which your classification model is confused when it makes predictions.

Predicted Label True Positive = 0.8516 False Negative = 0.1484

False Positive = 0.3765 True Negative = 0.6235

E. Long Short-Term Memory (LSTM):

Long Short-Term Memory (LSTM) networks are a kind of repeating brain network that are equipped for learning conditions in successions for expectation issues. They are valuable for spaces like machine interpretation, discourse acknowledgment, and more where it is essential to recollect past setting. It very well may be interesting to completely get a handle on what LSTMs are and the way in which terms like bidirectional and grouping to-succession connect with this field. LSTM networks are used to order and make estimates on words as per time series information, since there can be slacks in term between important occasions in a period series [2]. LSTMs were created to deal with the detonating and evaporating slope gives that can happen while preparing customary RNNs. An advantage of LSTMs over RNNs, stowed away Markov models, and other succession learning strategies and models is their general heartlessness in various applications.

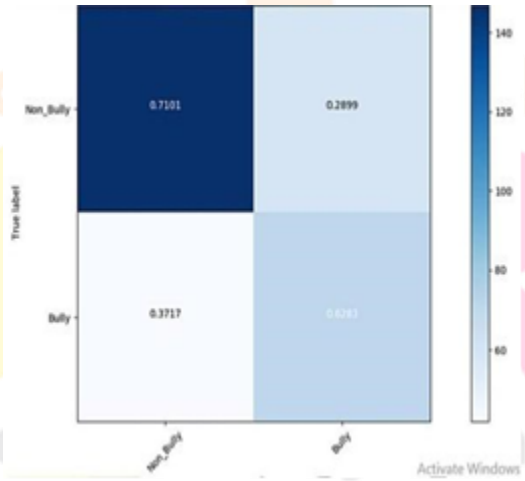


fig 2. the confusion matrix of lstm

F. Bidirectional LSTM:

Bidirectional LSTMs are an expansion of standard LSTMs that can further develop execution on arrangement characterization errands. In issues where the full info arrangement is accessible forthright, Bidirectional LSTMs train two LSTMs rather than one on the information succession. This gives extra setting to the organization and can bring about quicker and more far reaching learning. The methodology includes copying the principal intermittent layer so there are two equal layers. The first information succession is taken care of into one layer, while a switched duplicate is taken care of into the other. Utilizing the succession bidirectionally was at first roused in discourse acknowledgment, since there is proof that the setting of the whole expression is

utilized to decipher discourse, as opposed to only a basic one-way translation. While Bidirectional LSTMs may not appear to be legit for all expectation errands, they can give benefits through superior outcomes to spaces where the full setting is significant.

III. SCOPE

Cyberbullying, characterized as harassment through internet, abuse and intimidation, is impacting people of all ages and backgrounds. This has led to a surge in demand for technology and solutions that can efficiently detect and combat online bullying. Many organizations are using machine learning algorithms and other advanced technologies to automate the process of monitoring online behaviour in order to address the problem. The range of behaviours that may be classified as bullying is broad, as is the scope of cyberbullying detection [8]. This can range from overt acts of harassment and threats, to more covert actions like exclusion and spreading embarrassing or inappropriate material. The dynamic nature of online communication makes recognizing cyberbullying challenging, which is one of the main issues. Any cyberbullying detection technology must be developed with privacy in mind and used responsibly and morally to prevent harm. Educating people on how to spot and report bullying is a crucial component in detection. This may involve teaching kids and teens the importance of speaking out against bullying, as well as training parents, teachers and other adults. It is vital to modify and improve detection techniques to keep pace with developments as new social media platforms and messaging apps emerge and online interactions change. If we cooperate, we can make everyone's online experience more secure and respectful. In addition to identifying cyberbullying cases, the scope of detection also includes supporting those impacted by bullying. This can be done by using the machine learning algorithms and other advanced technologies to analyse large volumes of data and flag potential cyberbullying instances for further review by a human moderator. It also involves providing information on where to report cyberbullying and how to safeguard personal data and online accounts. Overall, the scope of cyberbullying detection is broad, and involves using a range of tools and strategies to identify and stop online bullying.

IV. PROPOSED SYSTEM:

a) Message feed:

The message feed from the dataset is the system's input. This is the starting point of the system's process and the input further word embedding.

b) Word Embedding:

The information from the message feed is implanted into mathematical structure for the contribution of the CNN. Each word is addressed by a genuine worth vector. The appropriated portrayal of words is advanced by the procedure of move learning.

c) Convolution Neural Network (CNN):

Neural network layers receive vectors formed by word embedding as input. The CNN's convolution layer is its fundamental layer. Its output is sent to the max pooling layer, which creates a fully connected network. The fully connected layer is followed by the Softmax algorithm to generate the output.

d) System Flow:

Long Short-Term Memory networks, or LSTMs, have been viewed as the best methodology. In numerous perspectives, LSTMs are better than RNN and conventional feed-forward brain organizations. Their capacity to hold specific examples in memory overstretched timeframes is the justification behind this. Long Short-Term Memory (LSTM) frameworks perform minor increments and augmentations to the information. Cell states are the strategy by means of which data moves in lengthy transient memory LSTMs. LSTMs can specifically review or fail to remember data as such. At some random cell state, there are three unmistakable conditions in the data. There are three separate conditions on the information at a given cell state.

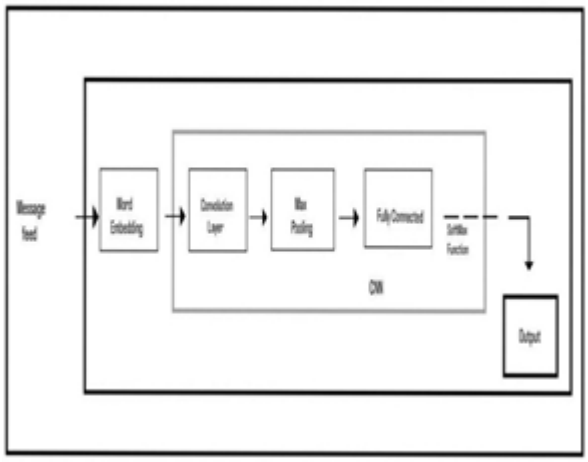


fig 3. block diagram of word detection

Long Short-Term Memory (LSTM) networks were proven very effective in sequence learning tasks. LSTMs can selectively remember or forget information over long time lags. This is done via cell states that flow information. An LSTM cell state depends on:

[1] The previous cell state (memory after prior time step)
[2] The previous hidden state (output of previous cell)
[3] The current time step input (new information)

LSTMs are analogous to conveyor belts in factories, shuttling information between layers. As data flows through, it may be added, modified, or removed like products on a belt.

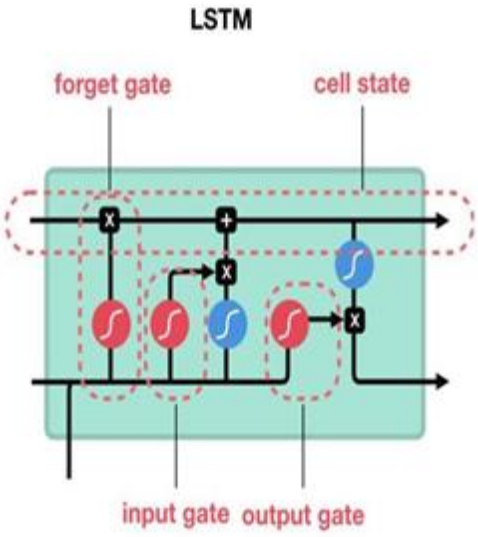


fig 4. lstm cells & their operations

A typical LSTM has blocks of memory called cells. Cell state and hidden state are two states whose information is passed to the next cell. Memory blocks are responsible for retrieving items, and three main mechanisms which are referred as gates, are used to manipulate that memory.

Three gates manipulate the cell's memory:

Forget Gate:

Decides what information to keep or forget.

Input Gate:

Updates cell states by filtering inputs between 0 and 1.

Output Gate:

Determines on what the following hidden state and results would be.

V. RESULT & ANALYSIS:

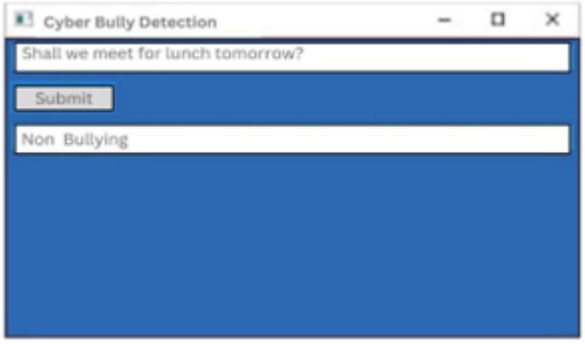
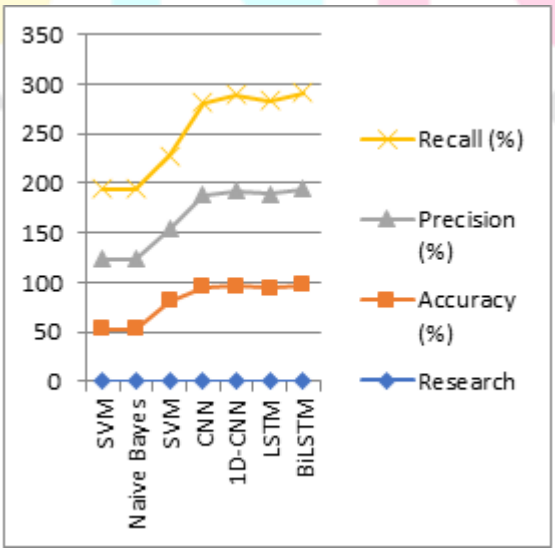


fig 5. non-bullying comment detected

fig 6. bullying comment detected

The techniques CNN and LSTM have the most accuracy percentage for cyberbully detection. This model is one of the reliable models to stop from cyberbullying and it detects when there are malice intent sentences or any threats.



graph-2: graphical representation of results

VI. CONCLUSION:

In conclusion, neural network-based cyberbully detection is a promising and noteworthy development in the ongoing effort to create more secure online spaces. The combination of state-of-the-art machine learning and artificial intelligence algorithms enables the proactive and automated identification and resolution of occurrences of online harassment.

REFERENCES

[4] Elaheh Raisi,Bert Huang "Weakly Supervised Cyberbullying Detection with Participant Vocabulary Consistency" Social Network Analysis and Mining, May 24,2018.

[5] Peng Zhou,Wei Shi,Jun Tian,Zhenyu Qi,Bingchen Li,Houng Wei,Hao,Bo Xu "Attention- based Bi-directional Long ShortTerm Memory Network for Relation Classification" proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,pages 207-212,August 12,2016.

[6] Alexis Conneau,Holger Schwenk,Yann Le cun "Very Deep CNN for Text Classification" Association for Computational Linguistics, Volume1, pages 1107-1116,7 April 2017.

[7] Elaheh Raisis,Bert Huang "Cyberbullying Detection with Weakly Supervised Machine Learning" International Conference on Advances in Social Networks Analysis and Mining IEEE/ACM,2017.

[8] Haipeng Zeng,Hammad Haleem,Xavier Plantaz,NanCao and Huamin Qu "CNN Comparator: Comparative Analytics of CNN" arXiv,15 Oct,2017.

[9] Vandana NandaKumar,Binsu C,Kovoor,Sreeja M.U "Cyber-Bullying Revelation in Twitter Data using NaiveBayes Classifier Algorithm" International Journal of Advanced Research in Computer Science. Volume 9, No. JanFeb 2018..

[10] Q. Li, proposed a new tweet sentiment classification approach using SSWE and WTFM produce classes based on the weighting scheme and text negation and a new text classification method.

[11] Roshni Jadhav, Grisha Chaudhari, Sumeet Rane "Cyber bulling Detection".

International Research Journal
IJNRD
Research Through Innovation