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. 14, Issue 1, January 2025

DOI: 10.17148/IJARCCE.2025.14135

Cyber Bullying Detection

Prof. Harshitha M[1], Karthik S M[2], Nandan Kumar[3], Pramod P[4], Yeshwanth M[5]

Assistant Professor, Information Science Department, East West Institute of Technology, Bangalore, India[1]

Student, Information Science Department, East West Institute of Technology, Bangalore, India[2]

Student, Information Science Department, East West Institute of Technology, Bangalore, India[3]

Student, Information Science Department, East West Institute of Technology, Bangalore, India[4]

Student, Information Science Department, East West Institute of Technology, Bangalore, India[5]

Abstract: Social media is a platform where many young people are getting bullied. As social networking sites are increasing, cyber bullying is increasing day by day. To identify word similarities in the tweets made by bullies and make use of machine learning and can develop an ML model automatically detect social media bullying actions. The objective of our project work is to show the implementation of NLP and LSTM which detects bullied tweets, posts, etc. A machine learning model is proposed to detect and prevent bullying on Twitter. Two classifiers i.e. NLP (Natural Language Processing) are used for identifying the complete sentence in the comments and LSTM (Long Short-Term Memory) for identification.

INTRODUCTION

Hate crimes are unfortunately nothing new in society. However, social media and other means of online communication have begun playing a larger role in hate crimes. For instance, suspects in several recent hate-related terror attacks had an extensive social media history of hate related posts, suggesting that social media contributes to their radicalization.

Detecting hate speech is a challenging task as First, there are disagreements in how hate speech should be defined. This means that some content can be considered hate speech to some and not to others, based on their respective definitions. Some recent studies claimed favourable results to detect automatic hate speech in the text. The proposed solutions employ the different feature engineering techniques and ML algorithms to classify content as hate speech. Regardless of this extensive amount of work, it remains difficult to compare the performance of these approaches to classify hate speech content. To the best of our knowledge, the existing studies lack the comparative analysis of different feature engineering techniques and ML algorithms.

To address this, we propose a system which employs Long-Short Term Memory, Natural Language Processing techniques and the classification is done using ensemble machine learning approach that incorporates various classification techniques.

PROPOSED SYSTEM

In this project, a solution is proposed to detect cyber bullying. The main difference with previous research is that we not only developed a machine learning model to detect cyber bullying content but also implemented it on particular locations real-time tweets using Twitter API. In Data Pre-processing, it is important to ensure that our dataset is good enough for analysis. This is where data cleaning becomes extremely vital. Data cleaning extensively deals with the process of detecting and correcting of data records. We had to make an intelligent decision regarding the type of feature that we want to select to go ahead with our machine learning model. In test train split we are splitting the dataset for training and testing for crating model and prediction. Then apply the algorithm for creating model for the sentiment classification.

METHODOLOGY

In This proposed model is a prototype for a Cyberbullying detection system which can be used for social media platforms for automated checking and control of Hate speech. The data for training is cleansed and pre-processed before being fed into stacked word embeddings. Then the CNN-Bi LSTM deep learning model is trained to perform better than regular deep learning models trained standalone. The model is saved for its use in the website. The website is similar to any social media platform where the user has access to many features. Admin will have privileges to view content status. Even though this work is a prototype, it is still a step towards getting a better result.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025
DOI: 10.17148/IJARCCE.2025.14135

• Dataset Analysis - The acquired labelled data in 3 languages, i.e., Hindi, English and Hinglish, from numerous open-source dataset sources go towards the text preprocessing stage which involves Data Cleaning, Data Integration, Data Transformation, Data Reduction and Data discretization.

• Data Cleaning - Any irrelevant attributes, empty cells and NaN values are removed. The data is also formatted so that the data type across the dataset is uniform.

• Data Transformation - As the three datasets are acquired from different sources, compiling them in their original form will not be compatible because of the difference in classification labels. To proceed with these datasets, it is important to get rid of different label sets and using one single classification technique, 0-1 classifier, which will tell us if the text contains content of Hate speech or not, thus making it a black and white area to train our model and eliminating any Gray possibilities.

• Data Integration - All the datasets are integrated to one csv file that is used for further text preprocessing.

• Data Discretization - In this stage the data was tokenized, i.e., splitted the sentence into words for easy evaluation of data.

• Data Reduction - In this text preprocessing stage, certain things are removed such URLs, special characters, '@' and stopped words from tweets and converted all the text into lower case. Further, stemming is performed, which is transforming a word to its root form, and lemmatization, which reduces the words to a word existing in the language. This stage helps in reducing data into its simplest possible form.

After the preprocessing of data is completed, we move towards the building and training of the model stage. As shown in Fig. 1, for building our CNN-BiLSTM model, Word Embedding approach is used as it solves various issues that the simple one-hot vector encodings have. Most crucial thing is that word embeddings boost generalization and performance. We will stack 2 word embeddings which are GloVe and FastText. A combination of embeddings has been established to produce the best results. After the stacking of word embedding, CNN-BiLSTM model is built. As a hybrid technique has shown the potential of reducing sentimental errors on increasingly complex data. An ensemble ML model is also built, in which feature extraction technique and unigram feature engineering are used. The proposed CNN-BiLSTM model is compared with an ensemble ML model to draw out a comparison on the accuracy.

CNN-BiLSTM Architecture

A single machine learning or deep learning model can predict the outcome rather accurately when applied to specific domains, but each has its own set of advantages and downsides. LSTM usually produces superior results, but it takes longer to process than CNN, and CNN has fewer hyperparameters and requires less supervision. In the meanwhile, the LSTM is more accurate for long sentences but takes longer to analyze. Because RNN has a major gradient loss issue when processing sequences, the perception of nodes in the front decreases as nodes get further back. To tackle the problem of gradient vanishing, BiLSTM is used. It solves the problem of fixed sequence to sequence prediction. RNN has a limitation where both input and output have the same size. So it fails in case of machine translation where input and output have different sizes or case of text summarization where input and output have a different length, which is not the case with BiLSTM. The concept of combining two (or more) methods is offered as a way of implementing the benefits of both while also addressing some of the drawbacks of existing techniques

.

A CNN BiLSTM is a bidirectional LSTM and CNN framework that is concatenated. It trains both character-level and word-level characteristics in the initial formulation for classification and prediction. The character-level properties are induced using the CNN layer. To derive a new feature vector using per-character feature vectors such as character embeddings and (preferably) character type, the model includes a convolution and a max pooling layer for each word.

Combining different variation yields multiple hybrid approaches that we have tested:
1. Glove+Fasttext⟶CNN⟶BiGRU⟶adam(dense,conv1d=relu;out=sigmoid),maxlen= 25
2. Glove+Fasttext⟶CNN⟶BiLSTM⟶adam(dense,conv1d=relu;out=sigmoid),maxlen=25
3. Glove+Fasttext⟶BiLSTM⟶BiGRU⟶adam(dense,conv1d=relu;out=sigmoid),maxlen=25
4. Glove+Fasttext⟶CNN⟶BiGRU⟶adam(dense,conv1d=relu;out=sigmoid),maxlen=25, trainable=True
5. Glove+Fasttext⟶BiLSTM⟶BiGRU⟶adam(dense,conv1d=relu;out=sigmoid),maxlen= =25,trainable=True
6. 25⟶Spatialdropout1D,GlobalMaxpooling1D,GlobalAveragePooling1D

The CNN-BiLSTM model that is to be used has the following features:

• Stacked Word Embedding: A distributed representation of words where different words that have a similar meaning (based on their usage) also have a similar representation. Two of such word embeddings are glove and fastext and stacking of these two embeddings provide better results

• Convolutional Model: A feature extraction model that learns to extract salient features from documents represented using a word embedding.

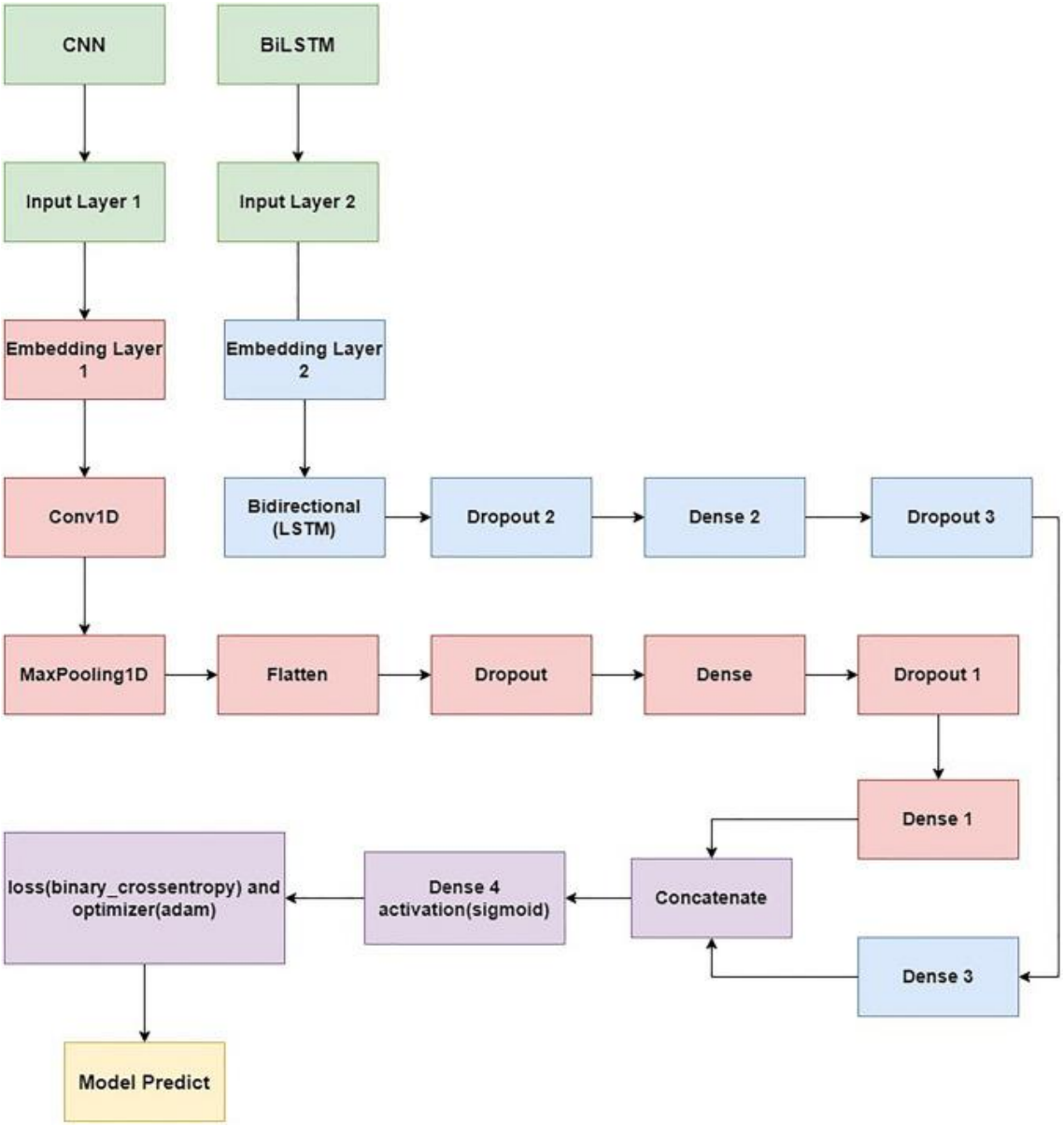
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025

DOI: 10.17148/IJARCCE.2025.14135

• Fully Connected Model: The interpretation of extracted features in terms of a predictive output.



Input layer t — The length of input sequences is defined by the input layer.
Embedding layer — 100-dimensional real-valued representations and an embedding layer set to the vocabulary's size.
Conv1D layer — Using 32 filters and a kernel size corresponding to the amount of words to read simultaneously.
MaxPooling1D — Merge the result of the convolutional layer with this layer.
Flatten layer — For concatenation and to convert the three-dimensional output to two-dimensional
Transfer function — Rectified Linear.
Kernel sizes— 3
Number of filters— 100
Dropout rate— 0.5
Weight regularization (L2) — 3
Batch Size — 128
Update Rule — Adam

The Adam optimizer is computationally more efficient, requires slight memory, is invariant to diagonal resizing of gradients, and it is well suited for problems with a lot of data/parameters. We will perform the best parameter using grid

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025
DOI: 10.17148/IJARCCE.2025.14135

search and 10-fold cross validation. Now, Convolutional Neural Network (CNN) models are built to classify encoded documents as either Hate speech or non-Hate speech. Now, the CNN model can be defined as follows.

- One Conv layer with 100 filters, kernel size 3, and relu activation function;
- One MaxPool layer with pool size = 2;
- One Dropout layer after flattened;
- Optimizer: Adam
- Loss function: binary cross-entropy (suited for binary classification problem)
- Dropout layers are used to solve the problem of overfitting and bring generalization into the model. As a result, in hidden layers, it's best to keep the dropout value near 0.5.

RESULTS

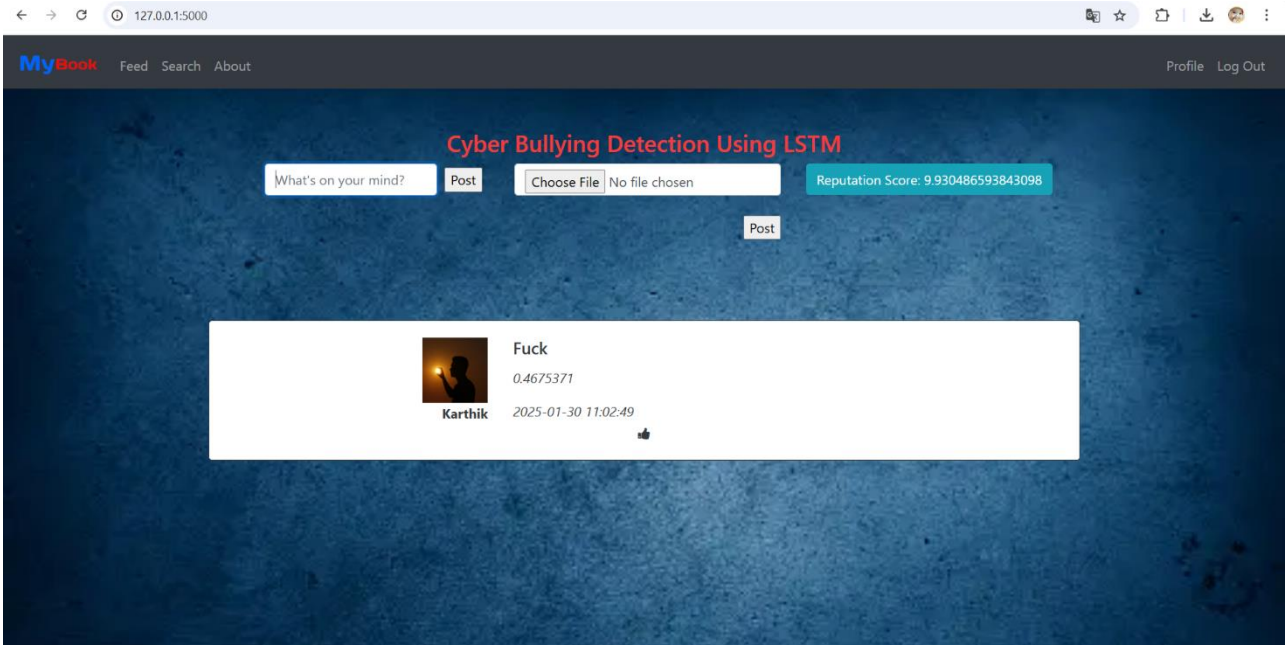


Fig. 1 Feed Interface

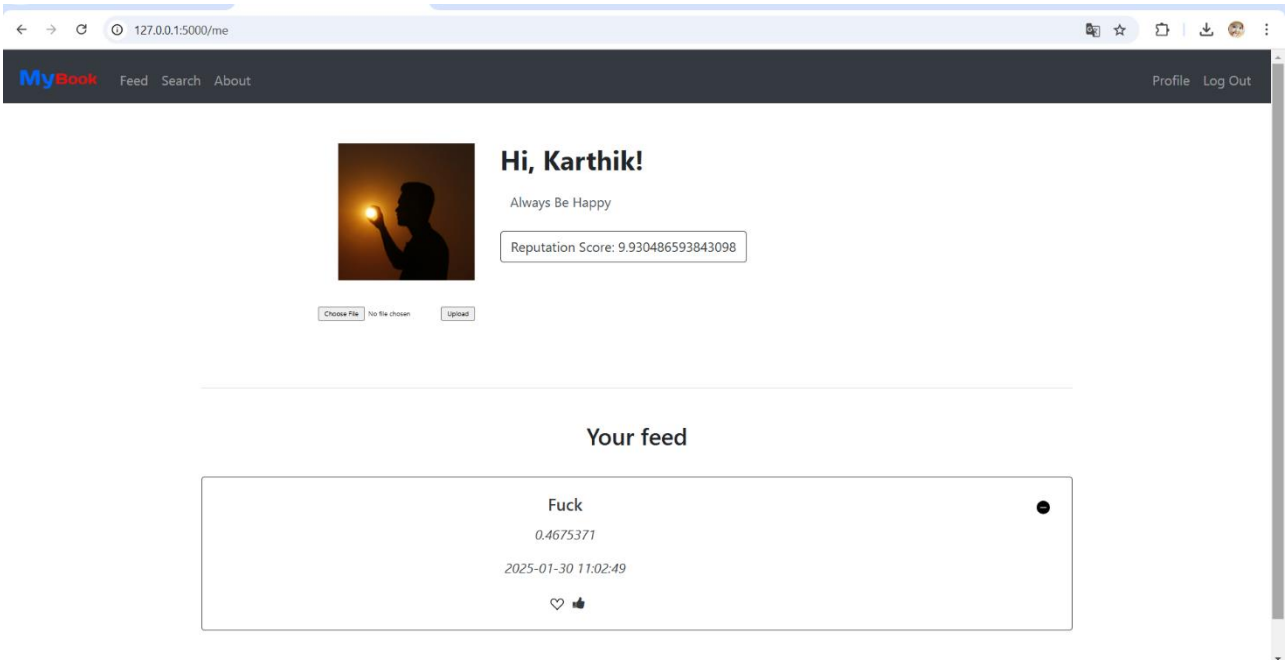


Fig. 2 User profile with reputation score

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025
DOI: 10.17148/IJARCCE.2025.14135

CONCLUSION

Internet crimes have become very dangerous because victims are continuously Being hunted, and there is little possibility of escape. Cyber bullying is one of the most critical internet crimes, and research has demonstrated its critical impact on the victims. The system uses accurate method of LSTM implementation using keras and helps in achieving precise results. This can help the users by preventing them for becoming victims to this harsh consequence of cyber bullying. Hence, compare to the existing model our technique is going to identify more accurate result of cyber bullying, where this new technique.

REFERENCES

[1] Das, Kumar & Garai, Buddhadeb & Das, Srijan & Patra, Braja. (2021). Profiling Hate Speech Spreaders on Twitter-Notebook for PAN at CLEF 2021.
[2] M. K. A. Aljero and N. Dimililer, "Genetic Programming Approach to Detect Hate Speech in Social Media," in IEEE Access, vol. 9, pp. 115115-115125, 2021, doi: 10.1109/ACCESS.2021.3104535.
[3] K. Sreelakshmi, B. Premjith, K.P. Soman, “Detection of Hate Speech Text in Hindi-English code mixed Data”, Procedia Computer Science, Vol. No. 171, Page No. 737-744, 2020.
[4] Al-Makhadmeh, Zafer, and Amr Tolba. "Automatic Cyberbullying detection using killer natural language processing optimizing ensemble deep learning approach." Computing 102, no. 2 (2020): 501-522.
[5] Ibrohim, Muhammad Okky, and Indra Budi. "Multi-label hate speech and abusive language detection in Indonesian twitter." In Proceedings of the Third Workshop on Abusive Language Online, pp. 46-57. 2019.
[6] Aditya Gayadhani, Vikrant Doma, Shrikant Kndre, Laxmi Bhagwat, “Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach”, IEEE International Advance Computing Conference(2018), 2018.
[7] Fauzi, M. Ali, and Anny Yuniarti. "Ensemble method for indonesian twitter Cyberbullying detection." Indonesian Journal of Electrical Engineering and Computer Science 11.1 (2018): 294-299.
[8] N. A. Setyadi, M. Nasrun and C. Setianingsih, "Text Analysis For Cyberbullying detection Using Backpropagation Neural Network," 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), 2018, pp. 159-165, doi: 10.1109/ICCEREC.2018.8712109.
[9] Kiilu, Kelvin & Okeyo, George & Rimiru, Richard & Ogada, Kennedy. (2018). “Using Naïve Bayes Algorithm in detection of Hate Tweets. International Journal of Scientific and Research Publications” (IJSRP). 8. 10.29322/IJSRP.8.3.2018.p7517.
[10] H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Cyberbullying detection," in IEEE Access, vol. 6, pp. 13825-13835, 2018, doi: 10.1109/ACCESS.2018.2806394.