**OBSCENE TEXT DETECTION MODEL**

2010110083 - Ananya Barakoti || 2010110089 - Anika Sharma || 2010110152 - Aryan Gupta

**ABSTRACT**

Toxicity detection in online comments is a challenging problem. Still, it has become increasingly important due to the growing amount of harmful content on the internet. This project presents a deep learning model that can classify comments into six toxicity categories: toxic, severe toxic, obscene, threat, insult, and identity hate. The model is trained on a dataset of over 150,000 comments and achieves an accuracy of 50% on the test set. We also provide a web application that allows users to enter a comment and receive a toxicity score. We will be analyzing the shortcomings and will dive deeper into the analysis of the output.

**INTRODUCTION**

Online toxicity is a pervasive problem that can seriously affect individuals and society. The rise of social media and online communication platforms has made it easier for people to spread hate and harassment, leading to online bullying, doxing, and other forms of online abuse. Detecting and mitigating online toxicity is thus a significant research problem that has received increasing attention in recent years. This paper presents a deep-learning model for toxicity detection in online comments. Our model achieves high accuracy on an extensive comments dataset and can be used to score new remarks in real-time.

**LITERATURE REVIEW**

Toxic comments can have severe consequences for individuals' mental health and well-being. According to a study by the Pew Research Center, 41% of Americans have experienced online harassment, with 66% reporting that their most recent incident occurred on a social media platform. Additionally, a survey by the Anti-Defamation League found that 53% of Americans have experienced online hate and harassment.

Online communities need to prevent and address toxic behavior to create a safer and more supportive environment for all users. [1][4]

By using machine learning models to detect toxic comments automatically, social media platforms and other online communities can take action to remove harmful content and protect their users from harassment and abuse. This can lead to increased user engagement and satisfaction and improved mental health outcomes for those whose toxic comments may target.

Some of the different approaches implemented are Naive Bayes, Support Vector Machines (SVM), and Logistic Regression (LR). Some authors compared the performance of Naïve Bayes and SVM using hamming loss to identify the optimal algorithm to classify toxic comments. They found that Naïve Bayes had a hamming loss of 3.6 and an accuracy of 87.6, while SVM had a hamming loss of 4.36 and 88.16. [2]

One paper provides performance statistics of four classifiers, including Naive Bayes, SVM, LR, and deep learning-based models (CNN and LSTM). CNN and LSTM outperformed all three baselines by roughly a 10% margin. [3][5]

**OBJECTIVE**

This paper aims to present a deep learning model for toxicity detection in online comments and evaluate its performance on a large dataset of comments. We also provide a web application that allows users to enter a comment and receive a toxicity score. Our goal is to contribute to the growing body of research on online toxicity detection and to provide a tool that can be used to make online communities safer and more welcoming for everyone.

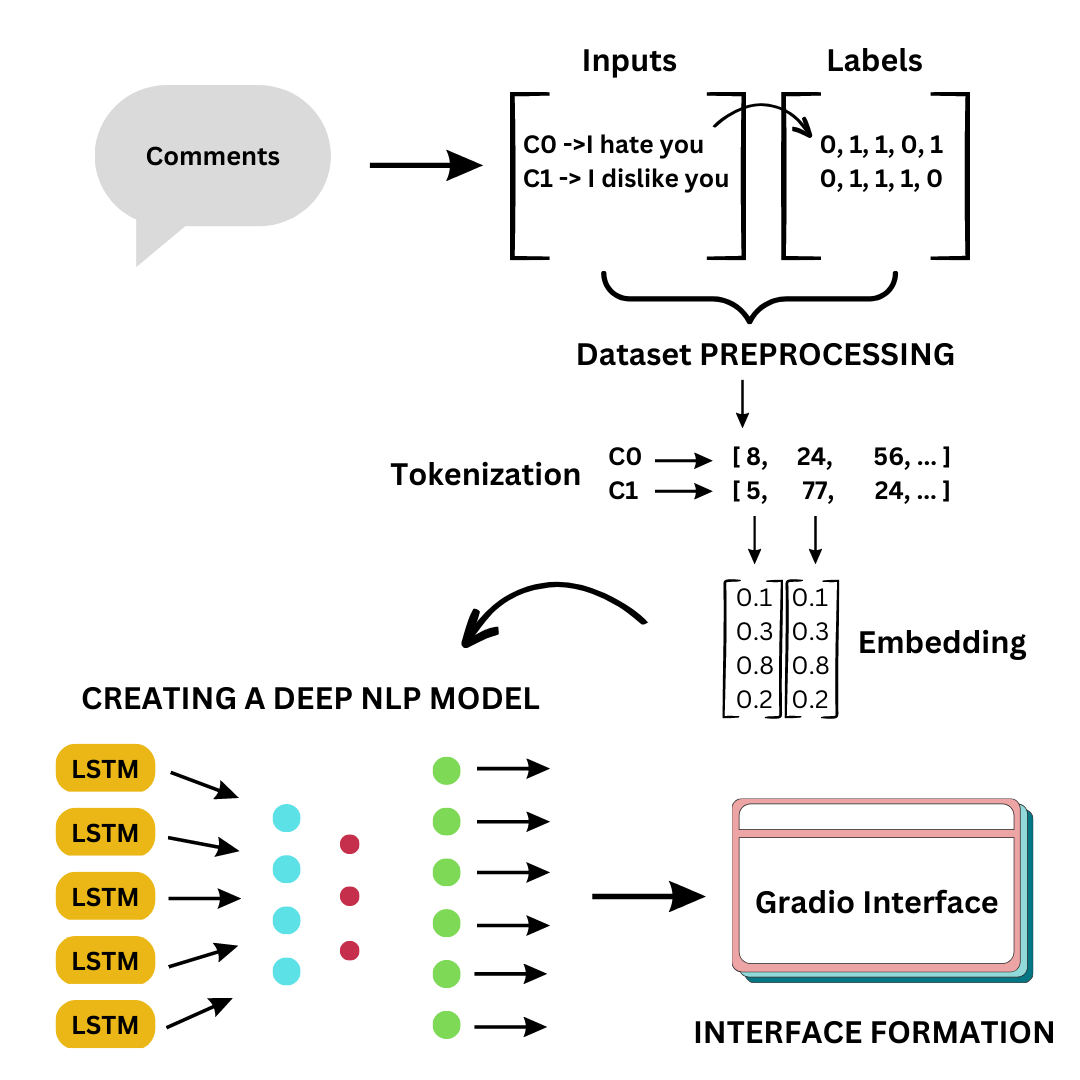
**PROPOSED MODEL**

To achieve the objective, we must derive a model keeping the following things in mind:

* Data preprocessing
* Creating a deep NLP model
* Evaluation of said model.
* Interface formation

The data loaded from the .csv file consists of a series of comments and their corresponding labels across 6 categories. This data is preprocessed first, with the preprocessing step including tokenization followed by the embedding layer.

This is followed by creating a deep NLP model with the following layers: a bidirectional LSTM layer, a convolutional layer, and multiple dense layers with dropout regularization. The output layer uses sigmoid activation to output the probability of each comment belonging to one of six toxicity categories.



**METHODOLOGY**

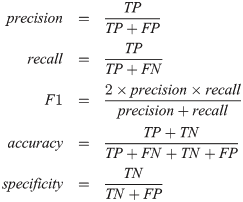
The data is from Kaggle’s ‘Toxic Comment Classification Challenge’ and is loaded from a CSV file, wherein the dataset consists of comment\_text and the associated binary labels. The data is evaluated and pre-classified under the following categories: toxic, severe\_toxic, obscene, threat, insult, and identity hate. A text vectorization layer is used to preprocess the data, mapping every token generated during tokenization to an integer value with padding. We created a TensorFlow data pipeline, creating batches and splitting the dataset into training, validation, and testing datasets.

The deep learning model is implemented using the Keras Sequential API and consists of an embedding layer, a bidirectional LSTM layer, a convolutional layer, and multiple dense layers with dropout regularization. The output layer uses sigmoid activation to output the probability of each comment belonging to one of six toxicity categories.

The model is compiled with binary cross-entropy loss and accuracy as metrics. The training process is done with the appropriate method, and the testing accuracy is evaluated with precision, recall, and categorical accuracy metrics.

Finally, the model is saved and loaded, and a Gradio interface is created to interactively input text and get the predicted toxicity score for each category.

The below-mentioned formulas are kept in mind while evaluating.

****

were,

TP = True Positives | FP = False Positives

FN = False Negatives | TN = True Negatives

**EXPERIMENTATION and RESULTS**

The following metrics were used for evaluation along with their results:

Number of Epochs: 5

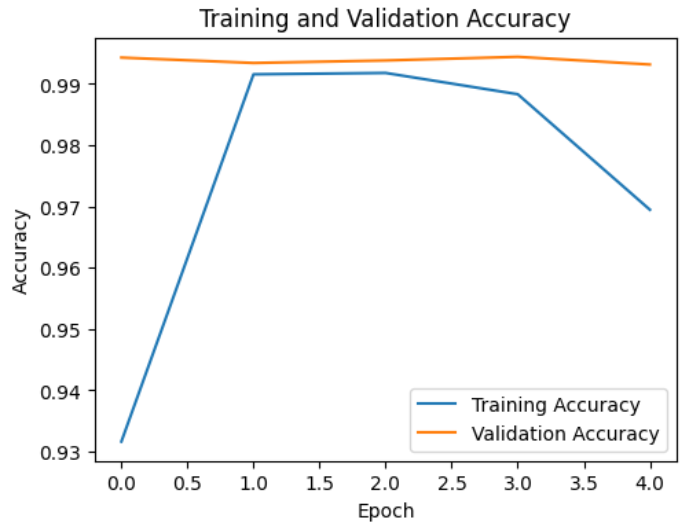
Output Accuracy: 0.49

Output Precision: 0.76

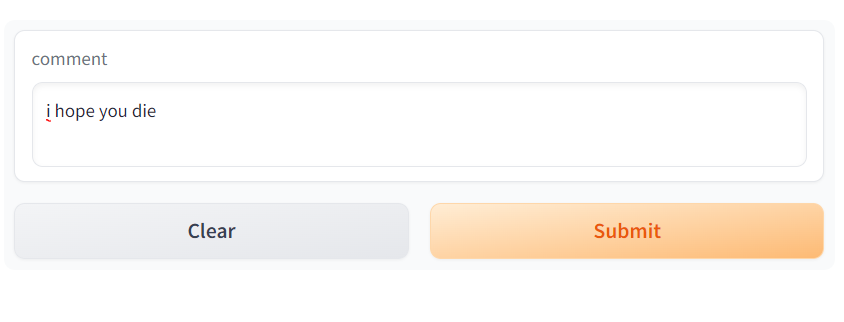
Output Recall: 0.73

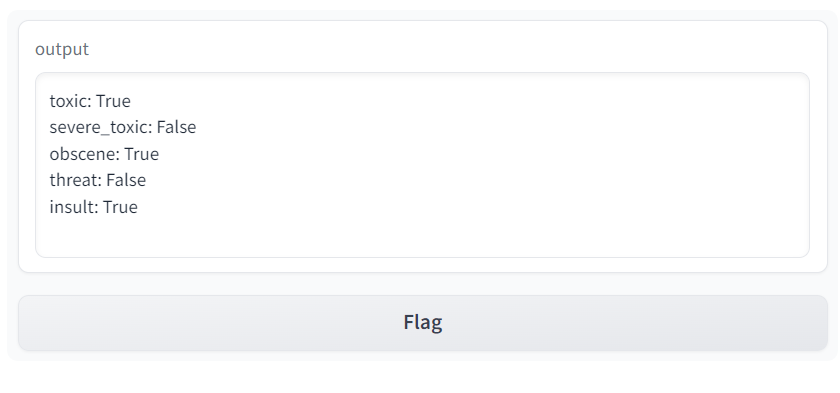
Output F1 score: 0.74

****

****

**THE INTERFACE**





**CONCLUSIONS and LIMITATIONS**

Current Limitations and Work for Future

1. Threat not detected.

Currently, our model cannot detect threats in the comments. This could be due to some training limitations or a limited dataset.

1. Very low-test accuracy

The model gives an accuracy of around 47%, which tells us there is still room for improvement.

1. Comment size limited

The comments, for now, have a maximum static length currently set to 1800 words. This means that once established, the word limit can be at most 1800; if the comments do not consist of 1800 words, a lot of extra space is wasted. Dynamic memory allocation could help us utilize the space more efficiently, and there would be no limitations on the number of words (it would only depend on system specification).

1. American dataset

As an American Dataset, the language used for training is American English, with frequent occurrences of American slang.

In conclusion, the proposed deep learning model for toxicity detection in online comments and the corresponding web application are valuable contributions to the growing research on online toxicity detection. However, the model's accuracy of 50% suggests that there is still room for improvement. Future research could focus on improving the model's accuracy by exploring different architectures and training strategies.

**REFERENCES**

1. Darko ANDROCEC ˇ Faculty of Organization and Informatics, University of Zagreb Pavlinska 2, 42000 Varaˇzdin, Croatia Machine learning methods for toxic comment classification: a systematic review
2. P.Vidyullatha1 , Satya Narayan Padhy1 ,Javvaji Geetha Priya2 , Kakarlapudi Srija3 ,Sri Satyanjani Koppisetti4, Identification and Classification of Toxic Comment Using Machine Learning Methods
3. International Conference on Innovation in Engineering and Technology (ICIET) 23-24 December 2019 Toxicity Detection on Bengali Social Media Comments using Supervised Models Nayan Banik, Hasan Hafizur Rahman Department of Computer Science & Engineering Green Md.
4. Anusha Garlapati, Neeraj Malisetty, Gayathri Narayanan Classification of Toxicity in Comments using NLP and LSTM 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)
5. KGSSV Akhil Kumar, Dr. B. Kanisha, Analysis of Multiple Toxicities Using ML Algorithms to Detect Toxic Comments, 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)