YouTube Pernicious Comments Bifurcation using Machine Learning

H K Anish Koushik, Nishanth Michael C, Chethan Ganiger
Student, Student, Student
Department of Information Science and Engineering,
Atria Institute of Technology, Bengaluru, India

Abstract – It is becoming more and more of a concern that the usage of online social media platforms has contributed to a sharp increase in damaging and derogatory remarks. It is imperative to deal with this problem and take action to lower toxicity. The classification of comment toxicity has significantly advanced recently, and numerous novel strategies have been put out. The goal of this analysis is to evaluate the type of statement and identify the various classes of toxic language, including obscene, identity-hateful, poisonous, insulting, and severely toxic language. Our system uses comments from websites like harmful or non-toxic as input. The goal of our model is to identify the toxicity class. Phased analysis is the goal of this research. In Phase I, the goal is to assess the toxicity of comments by providing data using multiple methodologies including TFIDF and spacy, which aid data in understanding how each word in a comment is categorised into a specific dangerous class. In this case, the algorithm will use test data comments to forecast the type of toxicity for test data, such as a toxic, threat, and so forth. Our system uses comments from websites like harmful or non-toxic as input. The goal of our model is to identify the toxicity class. Phased analysis is the goal of this research. In Phase I, the goal is to assess the toxicity of comments by providing data using multiple methodologies including TFIDF and spacy, which aid data in understanding how each word in a comment is categorised into a specific dangerous class. In this case, the algorithm will use test data comments to forecast the type of toxicity for test data, such as a toxic, threat, and so forth. Phase II involves data analysis to classify the comments as harmful or non-toxic. This encourages us to determine if a specific comment is harmful or not.

Index Terms – Term Frequency – Inverse Document Frequency (TF-IDF), Long-Short Term Memory (LSTM)

I. Introduction

People are using internet platforms more frequently, which allows them to communicate with one another by discussing opinions and sentiments about various events. This has helped to advance natural language processing (NLP). The abundance of internet comments serves as the backdrop for the problem statement. Users have had a very difficult time spotting poisonous comments. These internet discussions occasionally use language that can be classified into categories like toxic, insult, obscenity, severe toxic, threat, identity, and hate. Here, data must be classified using many labels, each with a binary categorization (0 and 1).

Therefore, it is believed that the issue statement is a multi-label classification problem. It's important to lessen unintentional prejudice in these online discussions. Even if social media provides the globe with a lot of good news, it also has drawbacks. In order to identify the toxicity of online comments, we will use Natural Language Processing along with deep neural networks in this study. In conjunction with recurrent neural networks having Long Short-Term Memory, word embeddings will be used (LSTM), separate quick texts to determine which model fits and performs the best.

Various reactions of opinion in the user comments when looking at the video content uploaded on YouTube are very affecting to the reputation of the video content and channel. The most important thing of information collecting is figuring out what other people do of thinking. Understanding something people think becomes the most important information for the content creator to make the acceptable video for the user. Therefore, an approach can be made to find out the perception of YouTube user on video content by using sentiment analysis data obtained from the textual content. Text mining approach becomes the best alternative to interpret the meaning of each comment. The classification of positive and negative content becomes very important for the YouTube user to assess how meaningful the content that has been published is based on user opinion comments.

II. LITERATURE SURVEY

Sl. No	Paper Title	Authors, Publisher & Publication Year	Problem Identified	Techniques used	Outcome
1	Analysis and Classification of User Comments on YouTube Videos	K.M. Kavitha, Asha Shetty, Bryan Abreo, Adline D'Souza, Akarsha Kondana Elsevier 2020	Users frequently remark negatively on the video contributors when they don't like the video that has been shared. In a situation with a large number of user views, the existence of a sizable number of strongly held opinions, whether favorable or negative, is likely to obscure the most insightful user comments that accurately reflect the video's substance. It is not as precise.	Using the video URL, the comments were extracted, and they were then manually divided into four classes. The retrieved remarks were automatically categorized using experiments. The precision (P), recall (R), and accuracy (A) metrics, estimated as follows, were used to evaluate the categorized comments.	According to how related they are to the video content as described in the description of the posted video, comments are divided into four categories: relevant, irrelevant, positive, and negative.

Sl. No	Paper Title	Authors, Publisher & Publication Year	Problem Identified	Techniques used	Outcome
2	Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – SupportVector Machine (NBSVM) Classifier	Abbi Nizar Muhammad, Saiful Bukhori, Priza Pandunata, IEEE 2019	The reputation of the video content and channel can be greatly impacted by the many user comments made in response to the uploaded YouTube videos. If there are more comments, it takes too much time to analyze each one, which is not practical.	While Nave Bayes is effective at classifying texts with modest numbers of data or document snippets, Support Vector is excellent at classifying texts with relatively high numbers of data or full-length documents. Together, Nave Bayes and Support Vector Machine provide increased performance and accuracy.	The application of the text mining idea along with the case folding, tokenization, filtering, and stemming phases produced the best results for word interpretation, increasing the accuracy of the classification model.

Sl. No	Paper Title	Authors, Publisher & Publication Year	Problem Identified	Techniques used	Outcome
3	Challenges for Toxic Comment Classification: An In-Depth Error Analysis	Betty van Aken, Julian Risch, Ralf Krestel, Alexander Löser Research Gate 2018	Found that the classification of harmful statements in online debates has to be improved and that there is a problem with misclassification. The difficulties include addressing the disparity in class, recognising and dealing with various forms of toxicity, and coping with sarcasm and irony in comments.	Employed error analysis to highlight the issues and constraints with the most up-to-date models for categorising poisonous remarks.	An in-depth exploration of the difficulties encountered in classifying poisonous comments in internet debates. The research exposes the shortcomings of the most recent state-of-the-art models and outlines a number of difficulties, such as class inequality, various forms of toxicity, sarcasm, and irony.
4	Identification and Classification of Toxic Comment Using Machine Learning Methods	P. Vidyullatha, Satya Narayan Padhy, Javvaji Geetha Priya Kakarlapudi Srija, Sri Satyanjani Koppisetti Turkish Journal of Computer and Mathematics Education 2021	Identification and classification of harmful comments on social media sites is a problem. the growing issue of cyberbullying and hate speech in online debates, as well as the requirement for automated methods to find and categorise such remarks. The difficulties of dealing with class imbalance, identifying and treating various forms of toxicity, and dealing with sarcasm and irony in comments, as well as the problems of constructing accurate and effective models for this purpose.	The approach for recognising and categorising harmful remarks is machine learning-based, and the paper suggests it. It is tested on a public ally available dataset.	The authors suggest a model that uses different feature extraction and text processing techniques to categorise comments as toxic or non-toxic. explains how to recognise and categorise harmful remarks using machine learning approaches, and it lays the groundwork for future study in this field.

Sl. No	Paper Title	Authors, Publisher & Publication Year	Problem Identified	Techniques used	Outcome
5	Classification of Online Toxic Comments Using the Logistic Regression and Neural Networks Models	Mujahed A. Saif, Alexander N. Medvedev, Maxim A. Medvedev, Todorka Atanasova International Conference on Machine Learning and Data Engineering (iCMLDE) 2018	Toxic comments in online debates are a problem, and automated techniques are needed to identify and categorise them. The paper emphasises the difficulties in detecting and dealing with various forms of toxicity, such as hate speech, cyberbullying, and offensive language, as well as the requirement for precise and useful models to handle this issue.	Identifies and categorises hazardous comments in online debates using a variety of machine learning techniques, including feature extraction, logistic regression, neural networks, and cross-validation.	Shows the efficiency of machine learning models for identifying and categorising harmful remarks in online debates and offers insightful information about the traits of toxic comments.

Sl. No	Paper Title	Authors, Publisher & Publication Year	Problem Identified	Techniques used	Outcome
6	A Machine Learning Approach to Comment Toxicity Classification	Navoneel Chakrabarty Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2019	Identifying and categorising poisonous remarks using machine learning techniques is a problem. presents a dataset of comments labelled with toxicity scores ranging from 0 (non-toxic) to 1 in order to highlight the prevalence and impact of toxic comments on online communities (toxic).	To classify comment toxicity, a combination of natural language processing methods, machine learning models, and evaluation measures was used.	Highlighted the value of applying natural language processing techniques and suitable evaluation measures for such projects by demonstrating the efficiency of machine learning approaches in tackling the challenge of comment toxicity classification.

III. PROPOSED SYSTEM

- Natural Language Processing is used in a novel way by the research and analysis to categorise the many types of toxicity in comments. Our system uses comments from websites like harmful or non-toxic as input. The goal of our model is to identify the toxicity class. Phased analysis is the goal of this research.

- The goal is to assess the toxicity of comments by providing information through a variety of ways that enable information to understand how each word in a comment is categorised into a certain harmful class. In this case, the algorithm will use test data comments to forecast the type of toxicity for test data, such as a toxic, threat, and so forth. To classify the comments into categories of toxic and non-toxic content, data is analysed. This advances us.

- Long-term dependencies are difficult for a standard RNN to learn, but they are learnable by the Long Short-Term Memory (LSTM) RNN. Similar to an RNN, an LSTM model features a chain-like structure, with each repeating structure's component known as an LSTM cell. An LSTM cell has three gates that regulate the flow of data into and out of the cell: an output gate, an input gate, and a forget gate. The data that will be deleted from the current cell by the forget gate is chosen. Following the input gate's determination of what fresh information will be delivered to alter the current state of the memory, the output gate determines what information exits the cell.

- By providing data through various methodologies, such as TFIDF and spacy, which aid data in understanding how each word in a comment is classified into a certain category of poisonous class, the goal is to analyse the toxicity in comments. In this case, the algorithm will use test data comments to forecast the type of toxicity for test data, such as a toxic, threat, and so forth. To classify the comments into categories of toxic and non-toxic content, data is analysed. This encourages us to determine if a specific comment is harmful or not.

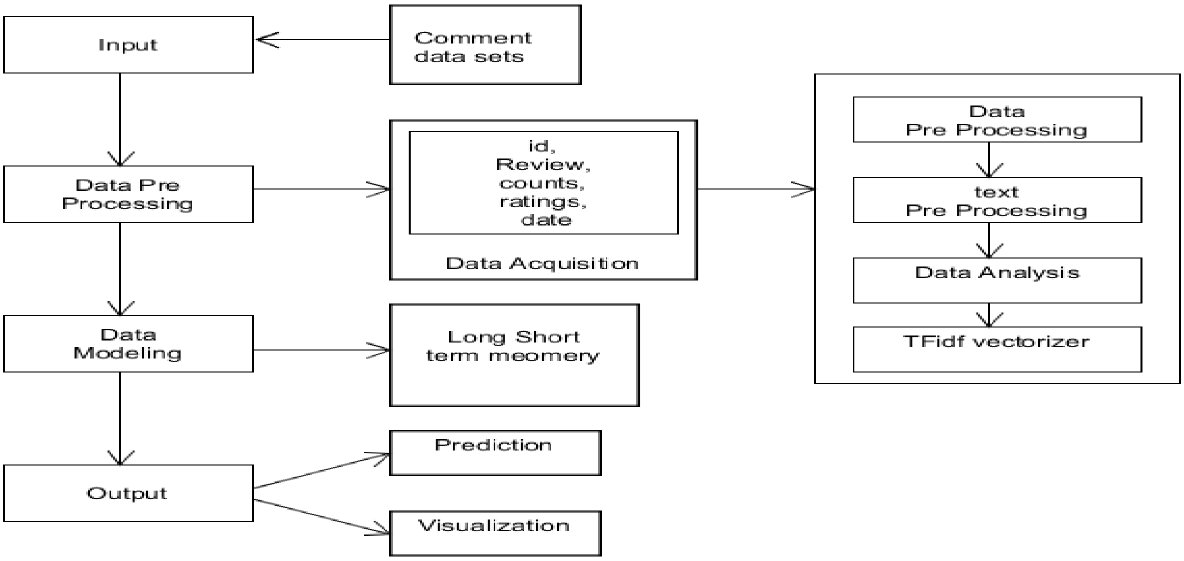


Fig: Architecture of the proposed system

The following items must be functional for this model:

1. Data is gathered from records of earlier Wikipedia analyses found in Kaggle datasets.
2. processing of raw data.
3. Preparing raw data for a machine learning model is a technique known as data pre-processing.
4. Using NLP
5. Ensure punctuation
6. Use of pauses
7. Stemming\sLemmatization
8. The divided train data from the pre-processing are fed into the LSTM algorithms after which the data is trained, and after that the test data is checked for accuracy.
9. The analysis aids in foreseeing the toxicity of comments.

- The product consists on of a model that functions based on:
- **Data Collection** - we are using the classic Comment Toxicity Datasets from Kaggle Repository.
- **Data pre-processing**- We'll frame the problem based on the dataset description and initial exploration. We are going to perform text pre-processing using NLTK library.
- **Data Analysis**- Carry our exploratory analysis to figure out the important features and creating new combination of features.
- **Data Preparation** - Using step 4, create a pipeline of tasks to transform the data to be loaded into our ML models.
- **Selecting and Training ML models** - Training a few models to evaluate their predictions using cross-validation.
- **Hyper parameter tuning** - For the models that produced encouraging results, fine-tune the hyper parameters.
- Utilizing the Flask web framework, deploy the model via a web service.

A data flow diagram (DFD) shows the "stream" of information passing through an information system graphically. The representation of data processing can also be done using an information stream diagram. It is customary for a developer to start by creating a setting level DFD that demonstrates how the framework and external pieces operate together. In order to exhibit greater information of the displayed framework, this setting level DFD is then "detonated."

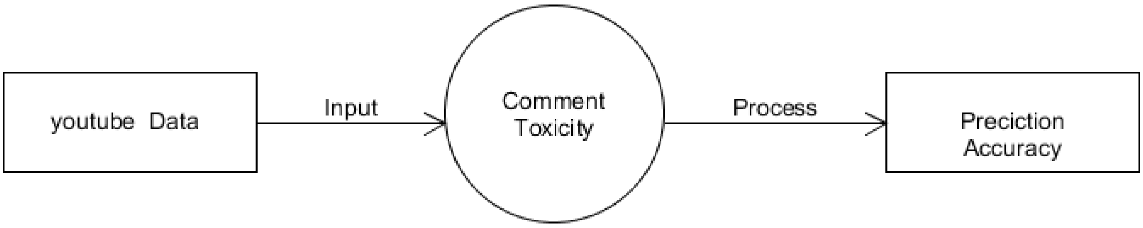


Fig: DFD 0

\

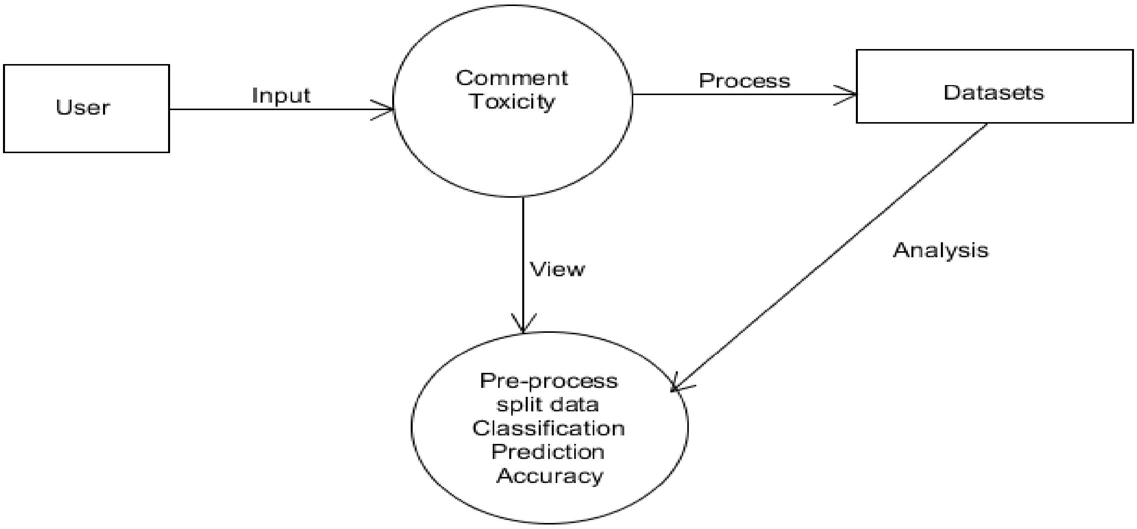


Fig: DFD-1

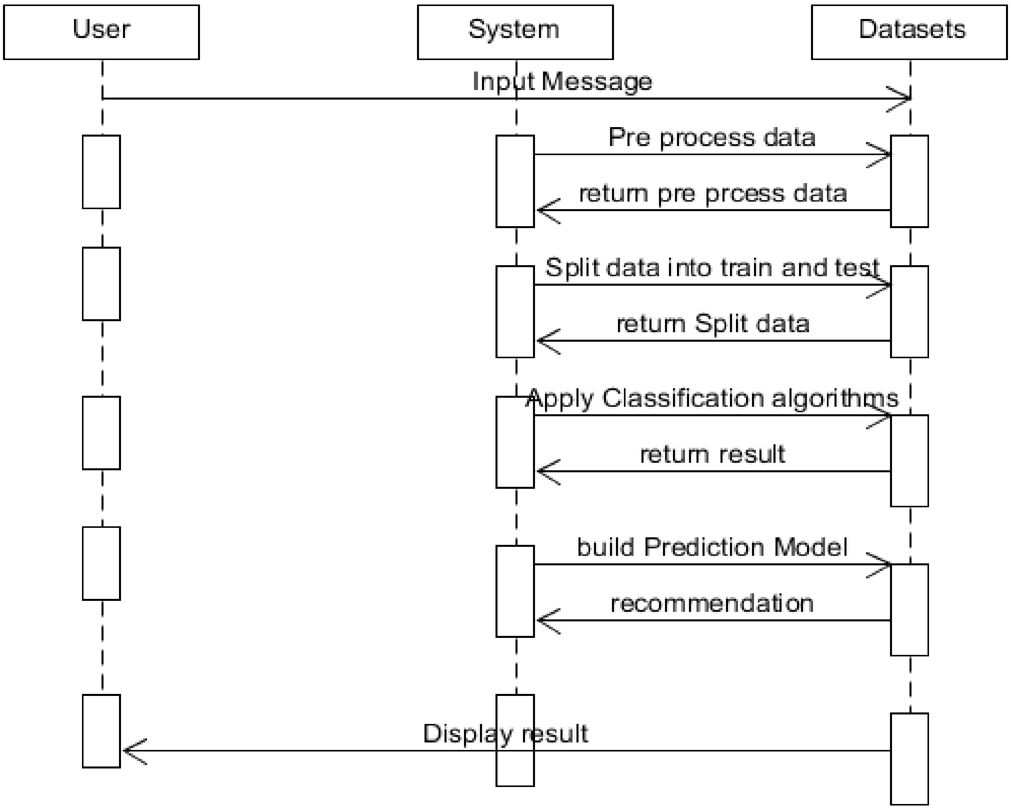


Fig: Sequence Diagram

- **Methodology:**

○ **Data Collection:** Data collection is one of the important things in our project. The right dataset must be provided to get robust results. We will be taking and analysing data from Kaggle. After seeing the accuracy, we will use the data in our model.

○ **Data Pre-Processing:** Making the data more machine readable is preferable because humans can grasp any sort of data but machines cannot. Raw data is frequently incoherent or lacking. Checking missing values, dividing the dataset, and other steps are all part of data pre-processing.

○ **Training Model:** Machines and models should learn by ingesting data, much like when feeding anything. The model will be trained using the retrieved data set. The input dataset for the training model is a raw set of data, while the desired output is a refined view offered from the same dataset. Different algorithms are used to refine the dataset and get the desired results.

○ **System Evaluation:** For the aforementioned project, we make use of a Kaggle dataset. However, this data set is in raw format. A collection of stock market valuation data for various companies makes up the data set. The first step is the creation of processed data from raw data. Feature extraction is used to achieve this since only a small number of the many attributes included in the raw data collected are necessary for the prediction. Feature extraction is a reduction operation. A structural model offers details on the composition, behaviour, and perspectives of a system.

○ **NLP STEPS:**

Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

Missing Data:

This issue develops when there are gaps in the data. There are several ways to handle it. Many of them include:

○ Ignore the tuples: This strategy only works when our dataset is sizable and numerous values are missing from a tuple.

○ Fill in the Missing Values: There are several methods to complete this operation. The missing values can be filled manually, by attribute mean, or by most likely value.

- **Algorithm:**

○ The LSTM (Long-Short Term Memory) algorithm, an RNN that can learn long-term dependencies, is employed in this model. A regular RNN finds it challenging to do this. An LSTM model, like an RNN, has a chain-like design, with each repeating structure's component being referred to as an LSTM cell.

○ Tokenization is the initial stage of this method. For each comment, this generates a series of integers. Since the length of each remark can vary, the tokenization process's output is padded to a set length of 200 characters. The Keras Embedding Layer, which learns embeddings, receives this next.

○ There was a 300 pixel embed size. An LSTM with 60 units receives the output of this embedding layer and outputs sequences. Next, data is sent to a pooling layer, a dense layer with 60 units, a dropout layer, and lastly a dense layer with 6 units and a sigmoid activation function. Six classes' probability are predicted by this.

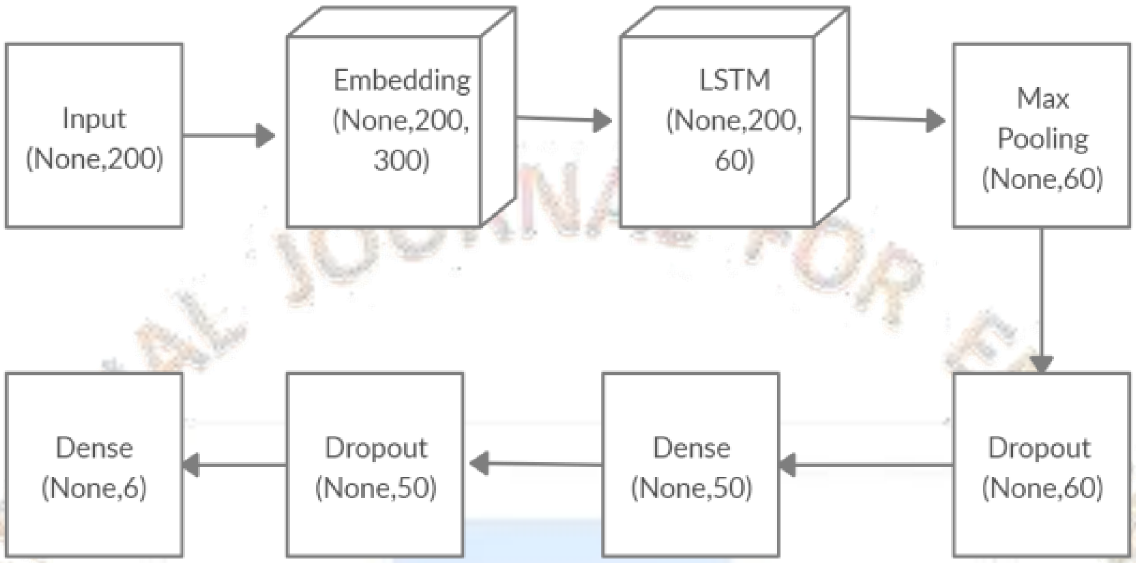
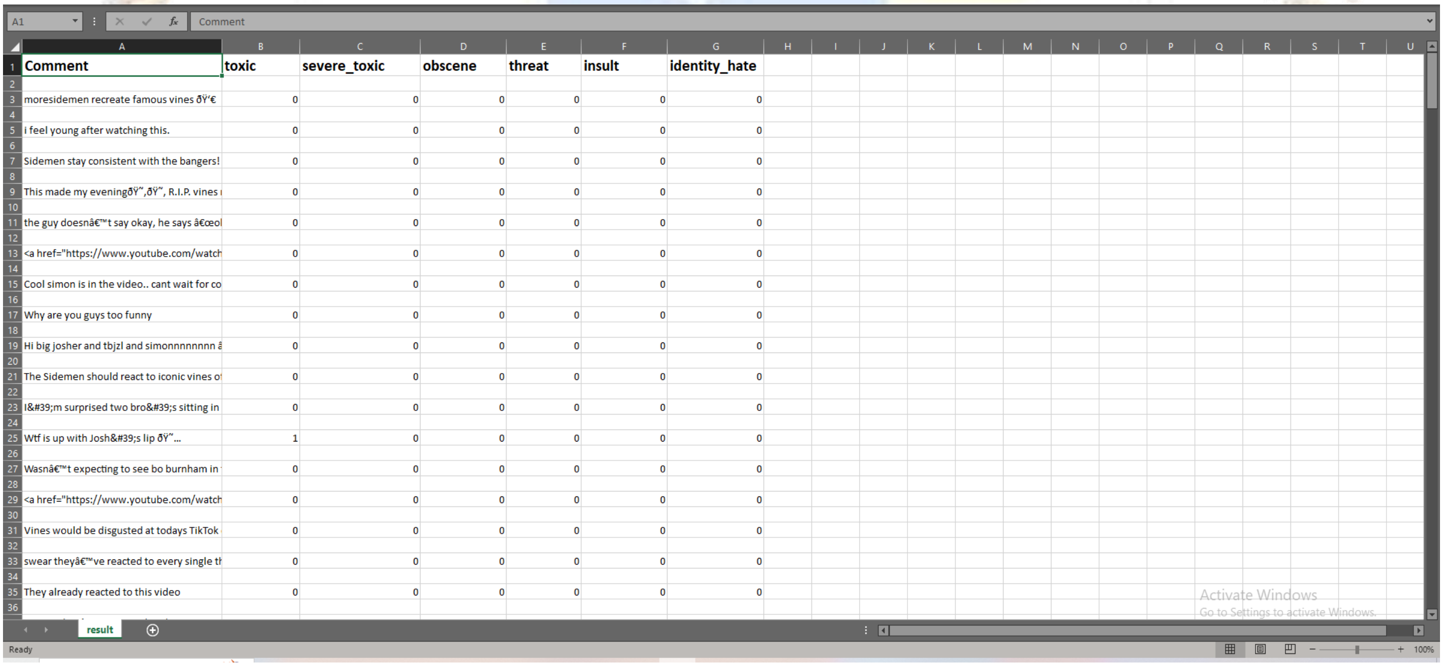


Fig: LSTM Architecture

IV. Results and Discussion:

The search for an effective model that can identify and forecast toxic remarks online has resulted in countless experiments and computations of the presence of toxicity of different kinds on online platforms, such as micro- and microblogging sites. This is significant in the field of research since user interaction on the internet is rapidly expanding. This research aims to identify the best ideal options for the further classification of harmful comments made online into six classifications. The obtained analysis shows that the LSTM (Long-Short Term Memory) perform better than the in terms of both the accuracy and time performance given the same number of epoch and are therefore preferable to use rather than with word-level embeddings. The analysis also compares the primary level neural network algorithms results with intricate Convolutional Neural Networks and Recurrent Neural Network compose: LSTM results. The proposed LSTM systems can be developed further, the word embeddings can be improved by employing more finely pre-processed data, and more accurate and promising results can be obtained.



Comment	toxic	severe_toxic	obscene	threat	insult	identity_hate
moresidemen recreate famous vines ðŸ‘€	0	0	0	0	0	0
i feel young after watching this.	0	0	0	0	0	0
Sidemen stay consistent with the bangers!	0	0	0	0	0	0
This made my eveningðŸ˜‚ðŸ˜‚, R.I.P. vines	0	0	0	0	0	0
the guy doesnâ€™t say okay, he says â€œol	0	0	0	0	0	0
<a href="https://www.youtube.com/watch	0	0	0	0	0	0
Cool simon is in the video.. cant wait for co	0	0	0	0	0	0
Why are you guys too funny	0	0	0	0	0	0
Hi big josher and tbjzl and simonnnnnnnn â	0	0	0	0	0	0
The Sidemen should react to iconic vines of	0	0	0	0	0	0
I&#39;m surprised two bro&#39;s sitting in	0	0	0	0	0	0
Wtf is up with Josh&#39;s lip ðŸ˜…	1	0	0	0	0	0
Wasnâ€™t expecting to see bo burnham in	0	0	0	0	0	0
<a href="https://www.youtube.com/watch	0	0	0	0	0	0
Vines would be disgusted at todays TikTok	0	0	0	0	0	0
swear theyâ€™ve reacted to every single th	0	0	0	0	0	0
They already reacted to this video	0	0	0	0	0	0

Fig: Results File

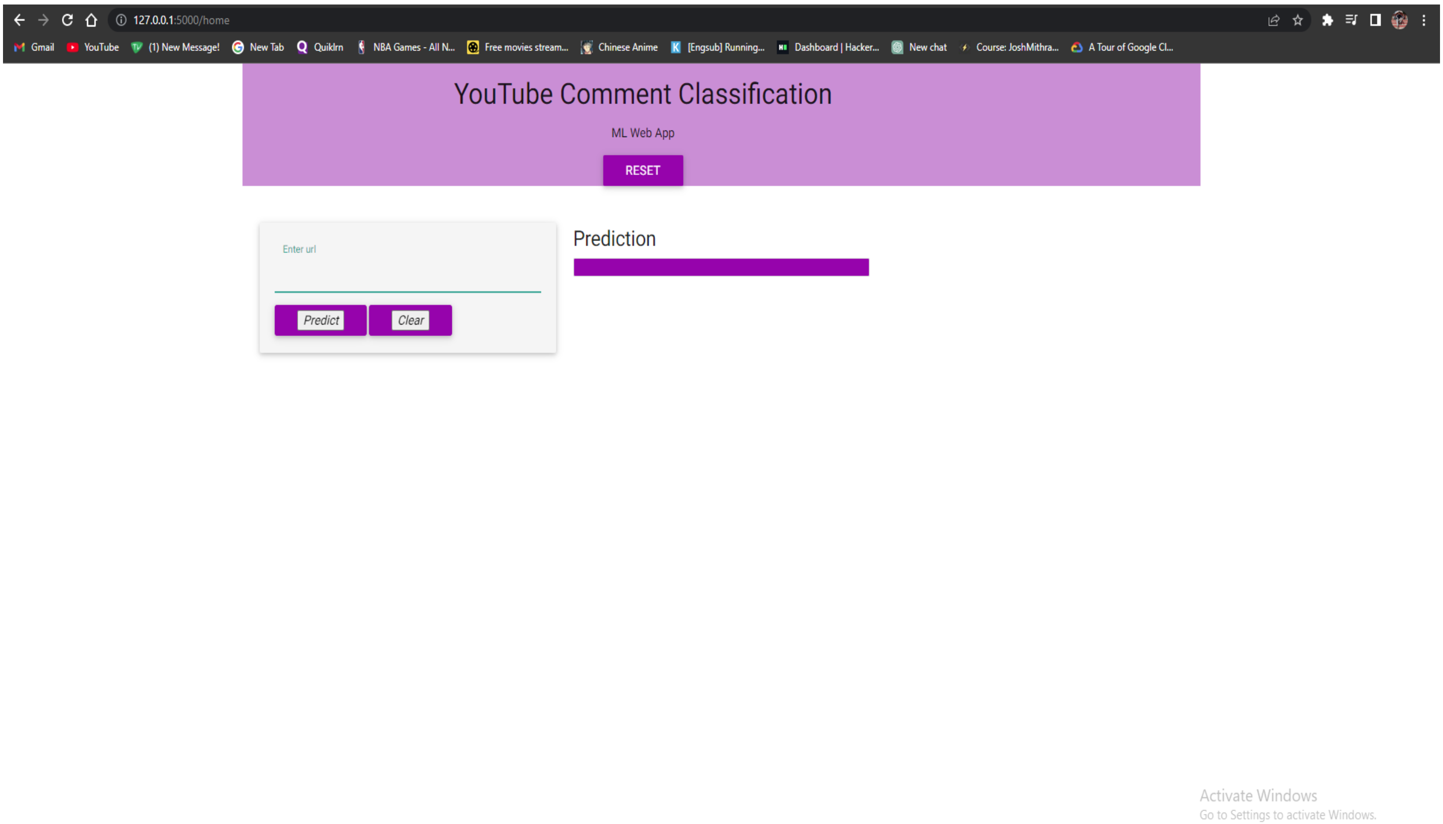


Fig: Front-End

V. REFERENCES

[1] K.M. Kavitha, Asha Shetty, Bryan Abreo, Adline D'Souza, Akarsha Kondana, "Analysis and Classification of User Comments on YouTube Videos", Elsevier, 2020

[2] Julio Savigny, Ayu Purwarianti, "Emotion Classification on Youtube Comments using Word Embedding", IEEE, 2017

[3] Abbi Nizar Muhammad, Saiful Bukhori, Priza Pandunata, "Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier", IEEE,2019

[4] L. Ceci. *YouTube Usage Penetration in the United States 2020, by Age Group*. Accessed: Nov. 1, 2021. [Online]. Available: https://www.statista.com/statistics/296227/us-youtube-reach-age-gender/

[5] J. Marsh, L. Law, J. Lahmar, D. Yamada-Rice, B. Parry, and F. Scott, *Social Media, Television and Children*. Shefeld, U.K.: Univ. Shefeld, 2019. [Online]. Available: https://www.stac-study.org/downloads/ STAC_Full_Report.pdf

[6] L. Ceci. *YouTubeStatistics & Facts*. Accessed: Sep. 01, 2021. [Online]. Available: https://www.statista.com/topics/2019/youtube/

[7] S. Alshamrani, A. Abusnaina, M. Abuhamad, D. Nyang, and D. Mohaisen, ``Hate, obscenity, and insults: Measuring the exposure of children to inappropriate comments in YouTube," in *Proc. Companion Proc. Web Conf.*, Apr. 2021, pp. 508515, doi: 10.1145/3442442. 3452314.