**Mini Project Report on**



**Toxicity Detection and Analysis in Social Media**



**Submitted in partial fulfillment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Student Name** **University Roll No.**

**Vinay Pandey 2019229**

***Under the Mentorship of***

**Dr. Ashwini Kumar Singh**

**Professor**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**Jan-2024**

GEU logo

**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Toxicity Detection and Analysis in Social Media”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Ashwini Kumar Singh, Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name : Vinay Pandey University Roll no: 2019229

**Table of Contents**

| **Chapter No.** | **Description** | **Page No.** |
| --- | --- | --- |
| Chapter 1 | Introduction | **1** |
| Chapter 2 | Literature Survey | **4** |
| Chapter 3 | Methodology | **6** |
| Chapter 4 | Result and Discussion | **10** |
| Chapter 5 | Conclusion and Future Work | **11** |
|  | References | **12** |

**Chapter 1**

**Introduction**

A concerning phenomena that has surfaced with the increased use of online platforms is toxic behavior on social media. It includes a variety of harmful behaviors and mindsets, such as hate speech, disseminating misleading information, and cyberbullying and harassment. This conduct frequently makes use of the anonymity offered by internet communications, enabling people to participate in destructive behaviors they may not otherwise. Beyond the negative effects on one's physical and mental health, toxic behavior also undermines the integrity of online communities. Toxic conduct must be identified and addressed in order to provide a healthy digital environment. We can all work to lessen the negative impacts of toxic conduct on social media by encouraging empathy, respect, and digital literacy.

* 1. **Risks of Toxic behavior in social media spaces**

Social media's ascent in popularity has completely changed how individuals interact, communicate, and exchange information. Even if these platforms are fantastic for networking and expressing oneself, they also give rise to a concerning phenomena called toxic behavior. On social media, behaviors that are damaging, unfavorable, and hurtful to people or the community at large are referred to as toxic conduct. These actions can take many different forms, undermining the original intent of these platforms by fostering a hostile online atmosphere.

Actions ranging from disseminating misleading information, hate speech, and trolling online to cyberbullying and harassment are all considered forms of toxic behavior. It frequently flourishes on the anonymity and distance that come with online connections, enabling people to indulge in activities they might not do in person. Beyond hurting a person's feelings, toxic behavior has an impact on mental health, creates division, and weakens the sense of community in online environments..

Toxic behavior must be recognized in order to create a better digital environment. It is critical to comprehend the various ways that toxicity manifests itself, from covert forms of online abuse to more overtones, such as subtle microaggressions. As social media continues to change, platform developers, users, and legislators must all work together to identify and mitigate harmful conduct. All people can have a more pleasant and inclusive online experience if we cultivate a culture of empathy, respect, and digital literacy.

In this project we are using a dataset of toxicity extracted from Kaggle which contains text id, the comments and various parameters of toxicity like age, race, gender color etc. which are used to tell us how toxic and on what grounds is the text toxic

**1.2 Objective**

Our project's goal is to provide a thorough toxicity analysis system for social networking sites. Our goal is to develop a powerful tool that can detect and evaluate hazardous content using cutting-edge natural language processing and machine learning approaches, in light of the widespread problem of harmful conduct online. Our goal is to differentiate between different types of toxic communication, such as hate speech, misinformation, and cyberbullying, by concentrating on the subtleties of online communication. The ultimate objective is to give users, content moderators, and social media platforms an efficient way to keep an eye on and reduce harmful conduct. With this initiative, we hope to improve the general well-being of social media users by promoting good digital interactions and safer, more inclusive online environments.

In this project, we will be utilizing a machine learning model, based on bidirectional Long Short Term Memory from the tensorflow library. Our dataset has been taken from kaggle which includes train and test splits with various parameters of toxicity like hate , racism, slur, severity etc.

All of this data has been classified using the NLP technique of vectorisation which transforms them into numeric tokens which can be fed to a dense Neural Network with a bidirectional LSTM layer to read the text through and through both backward and forward.

**1.3 Applications**

There is a great deal of promise for this toxicity study project to be used in solving the urgent issues that social media platforms are currently facing. Platforms can proactively improve content moderation efforts by putting in place a sophisticated system for detecting and evaluating hazardous content. With the use of this technology, social media managers will be able to quickly identify and take action against instances of hate speech, cyberbullying, and disinformation, fostering a more welcoming and safe online community. Furthermore, the toxicity study can yield valuable information that can be utilized to enhance platform policies, optimize user experience, and create focused interventions aimed at lessening the negative effects of toxic behavior. This project's application goes beyond platform management to encourage responsible digital citizenship by giving users the tools they need to keep an eye on and regulate their online activities.

**Chapter 2**

**Literature Survey**

# **Defining and detecting toxicity on social media: context and knowledge are key**

Publication Year: 2022

Author: Amit Sheth , Valerie L. Shalin , Ugur Kurşuncu et al..

Journal Name: Neurocomputing Volume 490

Summary:This Journal [1] delves into the complex issues surrounding the identification and definition of toxicity in social media. The writers recognize that online communication is always changing and stress the significance of taking knowledge and context into account when determining what constitutes harmful content. The study probably explores the challenges of using cutting-edge computational methods, such machine learning and natural language processing, to create efficient poison detection systems. The essay provides significant insights into the ongoing discussion on reducing toxic behavior on social media platforms by acknowledging the importance of information and context. It also suggests potential ways to improve the accuracy and context-awareness of toxicity analysis.

# **Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit**

Publication Year: 2021

Author: Hong Fan et al.

Journal Name: Machine Learning Technologies for Big Data Analytics

Summary: In this study[2], the authors tackle the issue of toxicity in social media. They employ a model based on Bidirectional Encoder Representations from Transformers (BERT) to efficiently detect and classify toxicity in user-generated content. The model is fine-tuned using a Kaggle public dataset focused on toxic comments and is subsequently tested on Twitter datasets related to the UK Brexit. The study demonstrates the effectiveness of their approach in identifying and analyzing toxic tweets, providing a promising tool for addressing harmful online content.

**An Automated Toxicity Classification on Social Media Using LSTM and Word Embedding**

Publication Year: 2022

Author: Ahmad Alsharef,Karan Aggarwal et al.

Journal Name: Journal of Computational Intelligence and Neuroscience

Summary: This study [3] focuses on the development of an automated system for toxicity classification in social media content. The study employs Long Short-Term Memory (LSTM) networks and word embedding techniques to enhance the accuracy of automated toxicity detection. LSTM networks are well-suited for capturing sequential dependencies in text, and word embedding helps represent words in a continuous vector space, improving the model's understanding of semantic relationships. By leveraging these technologies, the authors aim to create a robust tool for effectively classifying toxic content on social media platforms, contributing to the ongoing efforts to enhance online content moderation and foster a safer digital environment.

**Summary of Literature Survey:**

Here, I have reviewed the complex issues surrounding the identification and definition of toxicity on social media. The writers highlight how online communication is dynamic and how context and understanding are essential for spotting hazardous content. In order to create effective toxicity detection systems, the studies investigate cutting-edge computational techniques like machine learning, natural language processing, Bidirectional Encoder Representations from Transformers (BERT), and Long Short-Term Memory (LSTM) networks. The combined results provide insightful information to the ongoing discussion on decreasing toxic behavior online, highlighting the significance of context awareness and outlining viable paths for enhancing toxicity analysis accuracy.

**Chapter 3**

**Methodology**

**3.1 Tools and Technologies Used**

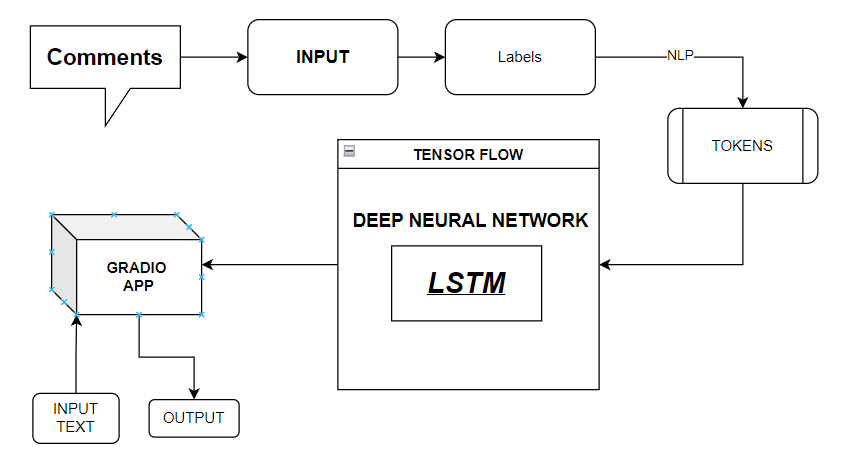
**Python :** Python is the language being used for this project. This is due to its many community-based characteristics, such as the fact that it has a wealth of potent tools available for scientific computing packages. Pandas and NumPy are two well-known and free software packages. These packages will drastically reduce and vary the amount of code required to create a certain program. Repetition becomes quick and effective as a result. The code written is shorter and easy to write. Hence Python was best suited for the project.

**Jupyter Notebook :** Users can create and share documents with live code, equations, visualizations, and narrative text using the interactive computing environment of Jupyter Notebook[5]. Its ability to handle many programming languages makes it a flexible tool for scientific research, machine learning, data analysis, and visualization. Users may describe their workflow, analyze data, and run code cells in real-time. Data scientists, academics, and educators choose Jupyter Notebook because it combines explanatory text with code to encourage repeatability and cooperation. It is an indispensable tool for data exploration and experimentation due to its intuitive interface and broad library support.

**TensorFlow -** TensorFlow [5] is a popular open-source framework for building and training machine learning models. It is useful for a wide range of tasks, from image and speech recognition to natural language processing and reinforcement learning Made by engineers and researchers acting on the Google Brain Team at intervals Google's Machine Intelligence analysis organization for the needs of conducting deep neural networks research and machine learning, but, the system is generally enough to be appropriate in a wide range of alternate domains as well.

Google Brain's second-generation system is TensorFlow. Whereas the reference implementation runs on single devices, TensorFlow can run on multiple GPUs and CPUs. TensorFlow is offered on Windows, macOS, 64-bit Linux and mobile computing platforms together with iOS and Android.

**3.2 Proposed Workflow**

****

**Figure 3.2.1** Proposed System Architecture and Workflow

The system presented here composes of six modules:-

1. Input as Dataset

2. Pre-processing

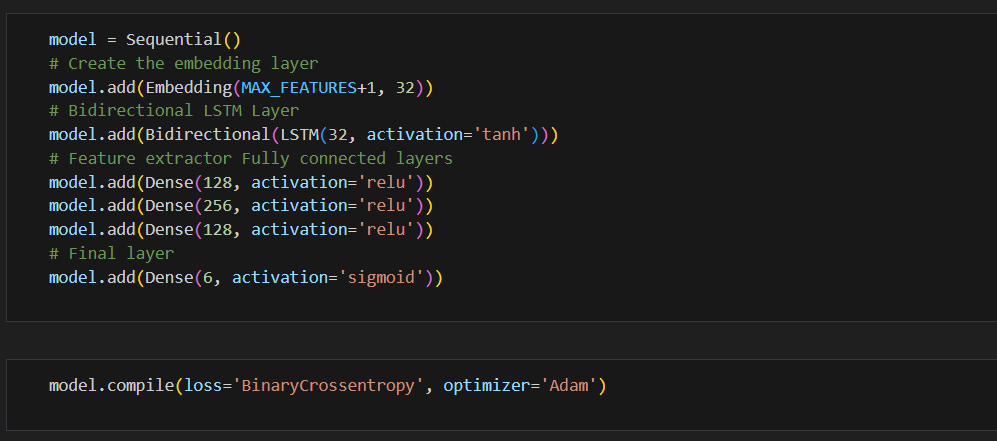
3. Data Labeling (Conversion to 0 to 1 float value)

4. Tokenize the text using built in Functions

4. Build & Model train a Deep Neural Network based on BiDirectional LSTM

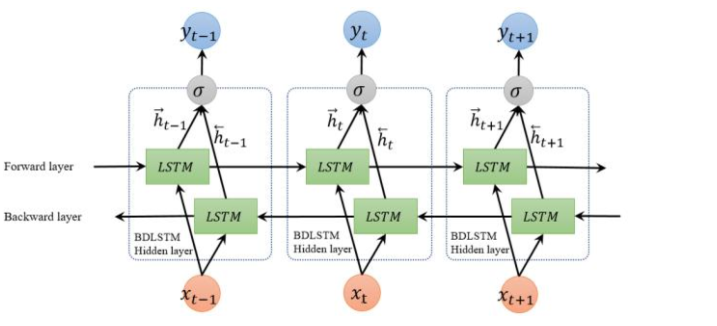
5. Take Input from gradio app and Output as Predicted Result

Data of the Toxic Comments and Labels have been taken from a website named Kaggle[6]. The data is then cleaned, scaled accordingly, and then split into values with range between 0 and 1, including float values which is then tokenized to be used in the model,values are then used to train our BiDirectional LSTM Model.



**Figure 3.2.2** Code for the Bi-Directional LSTM model with Deep Neural Network

**3.3 Bi-Directional LSTM Architecture**

****

**Figure 3.3.1** Bi-Directional LSTM Architecture

One kind of recurrent neural network (RNN) architecture called bidirectional long short-term memory (Bi-LSTM) is made to be able to receive and process sequential data in both directions. Conventional LSTMs only take into account prior knowledge and process input sequences in a unidirectional fashion. Bi-LSTMs, on the other hand, improve this ability by concurrently processing sequences in both the forward and backward directions.

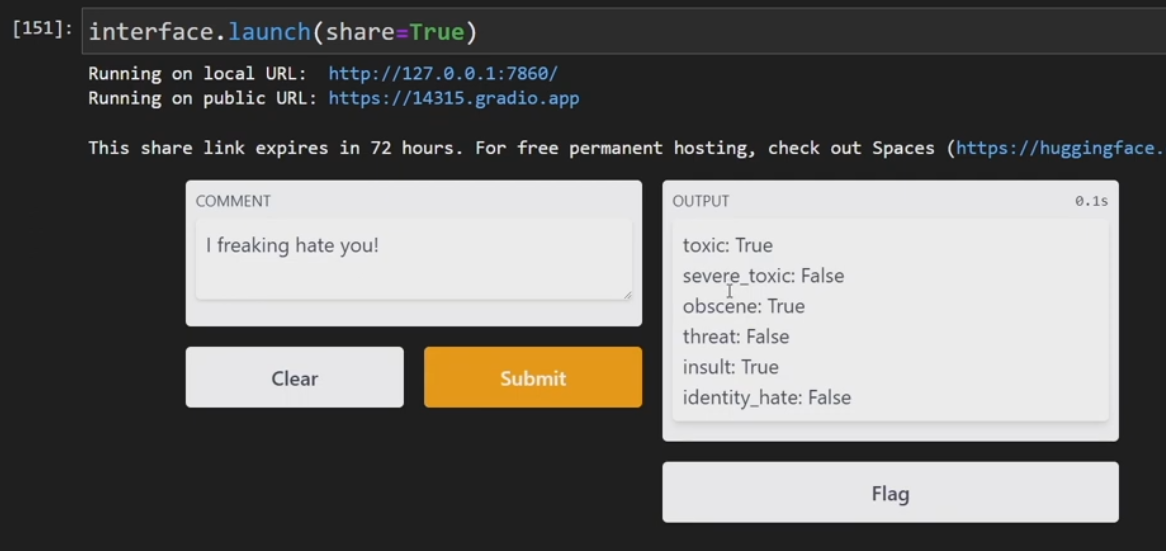
Two distinct hidden states are added to each time step in Bi-LSTMs, one capturing information from the past (ahead direction) and the other from the future (backward direction). This is the main characteristic of Bi-LSTMs. The model can gain a deeper comprehension of the context surrounding each piece in the sequence thanks to this bidirectional approach.

By utilizing information from both directions, Bi-LSTMs excel in capturing long-range dependencies in sequential data, making them particularly effective in tasks where context from both the past and future is crucial. This architecture is commonly employed in natural language processing tasks, such as sentiment analysis, named entity recognition, and sequence-to-sequence tasks, where understanding the context of words in a sentence is essential for accurate predictions.

**Chapter 4**

**Result and Discussion**

By following the described methodology, a TensorFlow Text Vectorizer was used to extract word weight using Natural Language Processing (NLP) whose output was used to train a Bi-directional Long Short Term Memory based Deep Neural Network which can tell with a good accuracy about whether a particular post is toxic or not and on what parameters. Since we used a bi directional model we were able to accurately study the given textUtilizing this model for testing on one case is given below



**Figure 4.1** Gradio app output after training the model and working on it-+

As we can see, the model is able to provide fairly accurate results, however , the model gives a false result at times. Hence the model, though well Trained still has few limitations and does not perform correctly in all conditions and could be made better by a larger dataset to analyze.

Training the model on more contextual parameters like date and source would improve the accuracy however the data classification in all the use cases would increase the load and complexity of the model considerably

**Chapter 5**

**Conclusion and Future Work**

Through this project, I have developed a system that can predict if a social media post is likely to be toxic or not, and on various grounds of toxicity, fairly accurately based solely on the content information . It has the potential to be applied to various post sentiment and social media platforms to ensure that the user generated content is not hateful or toxic towards any race, ethnic backgrounds or groups.

However it is important to note that the model , though provides a good accuracy over the original dataset, is still not a totally accurate analysis of the independent posts and needs more data to be trained on with several more parameters. A button or option to report false claims should be recommended with a deployment of this model

Although numerous machine learning models like BERT with higher accuracy are available, this model can be further enhanced by giving it a proper graphical interface, as well as training it with increased parameters like origin of post, context info, metadata, associated posts and groups. A better GUI could also be made with lightweight frameworks like Flask or Django which could make it easier to use. An API for easy integration into apps could also be made to benefit the user and make it easier for developers to integrate this system into their applications.

Moreover, there are a number of intriguing directions to pursue for the project's future development. Toxicological identification is a persistent difficulty due to the dynamic nature of online communication and the constantly changing social media ecosystem. Exploring how to use advances in deep learning and natural language processing to handle new types of harmful conduct could be the main focus of research efforts. Furthermore, adding user feedback mechanisms and real-time monitoring capabilities could improve the system's responsiveness and adaptability. Working together with social media companies to apply and improve toxicity detection algorithms in practical contexts may yield insightful information and help develop moderation tools that are more potent.

**References**

[1] Amit Sheth, Valerie L. Shalin, Ugur Kurşuncu, et al., "Defining and detecting toxicity on social media: context and knowledge are key," Neurocomputing, Volume 490, 2022.

[2] Hong Fan et al., "Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit," Machine Learning Technologies for Big Data Analytics, 2021.

[3] Ahmad Alsharef, Karan Aggarwal et al., "An Automated Toxicity Classification on Social Media Using LSTM and Word Embedding," Journal of Computational Intelligence and Neuroscience, 2022.

[4] Project Jupyter.” https://www.jupyter.org (accessed December. 25, 2023).

[5]TensorFlow : tensor flow 2.14.0 documentation.” https://www.tensorflow.org/ (accessed October. 29, 2023).

[6]Kaggle Competition : “ https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge” (Accessed October, 15, 2023)