Report On

Toxic Chat Detection

Submitted in partial fulfillment of the requirements of the Course project in Semester VIII of Final Year Computer Engineering

by
Ved Kokane(Roll No. 58)
Vikas Jamge (Roll No. 56)
Bhumit Malvi(Roll No. 62)

Mentor
Dr. Tatwadarshi P. N.

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

(A.Y. 2021-22)

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

CERTIFICATE

This is to certify that the Mini Project entitled **“Toxic Chat Detection”** is a bonafide work of **Ved Kokane(Roll No. 58), Vikas Jamge(Roll No. 56), and Bhumit Malvi(Roll No. 62)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **“Bachelor of Engineering”** in Semester III of Second Year **“Computer Engineering” .**

Dr. Tatwadarshi P. N.
Mentor

Dr Megha Trivedi
Head of Department

Dr. H.V. Vankudre
Principal

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Course Project Approval

This Mini Project entitled "Toxic Chat Detection**"** by **NVed Kokane(Roll No. 58), Vikas Jamge(Roll No. 56), and Bhumit Malvi(Roll No. 62)** is approved for the degree of **Bachelor of Engineering** in in Semester III of Second Year **Computer Engineering .**

Examiners

1……………………………………..
(Internal Examiner Name & Sign)

2………………………………………
(External Examiner name & Sign)

Date:

Place:

Contents

Abstract

Natural language processing, or NLP, is a type of artificial intelligence that deals with analyzing, understanding, and generating natural human languages so that computers can process written and spoken human language without using computer-driven language. Natural language processing, sometimes also called “computational linguistics,” uses both semantics and syntax to help computers understand how humans talk or write and how to derive meaning from what they say. This field combines the power of artificial intelligence and computer programming into an understanding so powerful that programs can even translate one language into another reasonably accurately. This field also includes voice recognition, the ability of a computer to understand what you say well enough to respond appropriately.

Acknowledgements

It is said that “learning is a never ending process.” While working on the project we have undergone the same experience of learning new things as we proceeded in our goal of building a Toxic Chat Detection System.

Working on the project was a new experience for us. As it opened a new gateway wherein we had an opportunity to work on a totally new concept as far as the engineering syllabus is concerned where most of the concepts are to be learned by rote.

The joy of working in a new domain and learning new things was a welcome experience for the four of us and all we have to say is that we have cherished all the moments as they came by, right from working on a project to making this report.

We would like to thank our Principal **Dr. Harish Vankudre** for constant motivation and support to excel and having faith in our ability. We would also like to thank our professor **Dr. Megha Trivedi** (Head - Department of Computer Engineering) for providing her views on the subject.

We would like to thank **Dr. Tatwadarshi P. N.,** who guided us and shared their knowledge & invaluable experience about the topic and gave their precious time towards solving our difficulties. We would also like to thank our college management for providing us with the facilities and infrastructure for working on the project.

____
Ved Kokane (58)

____
Vikas Jamge (56)

____
Bhumit Malvi (62)

Date :

1. Introduction

1.1 Introduction

Comment sections of online news platforms are an essential space to express opinions and discuss political topics. In contrast to other online posts, news discussions are related to particular news articles, comments refer to each other, and individual conversations emerge. However, the misuse by spammers, haters, and trolls makes costly content moderation necessary. Sentiment analysis can not only support moderation but also help to understand the dynamics of online discussions. A subtask of content moderation is the identification of toxic comments. To this end, we describe the concept of toxicity and characterize its subclasses. Further, we present various deep learning approaches, including datasets and architectures, tailored to sentiment analysis in online discussions. One way to make these approaches more comprehensible and trustworthy is fine-grained instead of binary comment classification. On the downside, more classes require more training data. Therefore, we propose to augment training data by using transfer learning. We discuss real-world applications, such as semi-automated comment moderation and troll detection. Finally, we outline future challenges and current limitations in the light of most recent research publications.

Posting comments in online discussions has become an important way to exercise one's right to freedom of expression on the web. These essential rights however are under attack: malicious users hinder otherwise respectful discus-sions with their toxic comments. A toxic comment is defined as a rude, dis-respectful, or unreasonable comment that is likely to make other users leave discussion. A subtask of sentiment analysis is toxic comment classification.In the following, we introduce a fine-grained classification scheme for toxic comments and motivate the task of detecting toxic comments in online discussions

Social media, blogs, and online news platforms nowadays allow any web user to share his or her opinion on arbitrary content with a broad audience. The Media business and journalists adapted to this development by introducing comment sections on their news platforms. With more and more political campaigning or even agitation being distributed over the Internet, seriousand safe platforms to discuss political topics and news in general are increasingly important. Readers' and writers' motivations for the usage of news comments have been subject to research. Writers' motivations are very heterogeneous and range from expressing an opinion, asking questions, and correcting factual errors, to misinformation with the intent to see the reaction of the community.

1.2 Problem Statement and Objectives

Toxicity has become prevalent in Today's Internet. Unhealthy comments everywhere make the environment unfit for healthy talks or discussions. To counter this problem we develop a Toxic Chat Detection system using Natural language processing.

Objectives

- To make a system which identifies whether the comment is Toxic or Not
- If Toxic identify the level of Toxicity

1.3 Scope

Much of the current work is focused in two major directions:

- To implement NLP preprocessing to identify text
- To use the LSTM network to train and identify toxic chat.

This project will make the Toxic Chat detection easier. Users can use this system to detect and possibly remove toxic comments.

1.4 Course Project Contribution

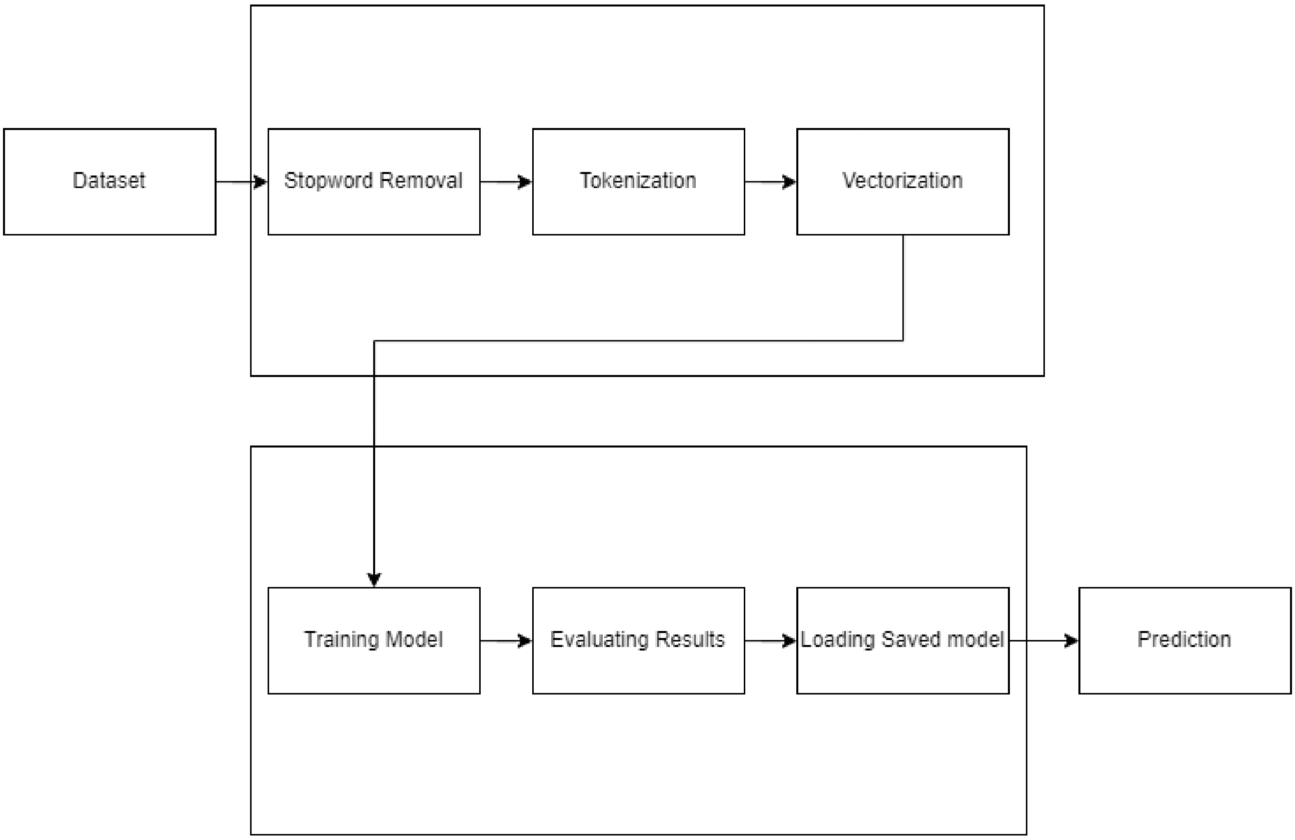
- Project Member Ved Kokane has implemented the LSTM Model required for the project
- Project Member Vikas Jamge has implemented the NLP preprocessing techniques
- Project Member Bhumit Malvi has Prepared and cleaned the Dataset.

2. Proposed System

2.1 Introduction

Comment sections of online news platforms are an essential space to express opinions and discuss political topics. In contrast to other online posts, news discussions are related to particular news articles, comments refer to each other, and individual conversations emerge. However, the misuse by spammers, haters, and trolls makes costly content moderation necessary. Sentiment analysis can not only support moderation but also help to understand the dynamics of online discussions. A subtask of content moderation is the identification of toxic comments. To this end, we describe the concept of toxicity and characterize its subclasses. Further, we present various deep learning approaches, including datasets and architectures, tailored to sentiment analysis in online discussions. One way to make these approaches more comprehensible and trustworthy is fine-grained instead of binary comment classification. On the downside, more classes require more training data. Therefore, we propose to augment training data by using transfer learning. We discuss real-world applications, such as semi-automated comment moderation and troll detection. Finally, we outline future challenges and current limitations in the light of most recent research publications.

2.2 Block Diagram



2.3 Algorithm and Process Design

- The Dataset is cleaned and Loaded in the csv file
- The data is preprocessed by using Stopword Removal from NLTK English stopword repository.
- The data is then tokenized using the same library.
- The tokenized data is then converted to array form using Vectorization algorithms.
- The array is then passed to a LSTM model to train.
- The trained Model is evaluated and Loaded.
- The test Data is predicted on the loaded Model

2.4 DETAILS OF HARDWARE AND SOFTWARE

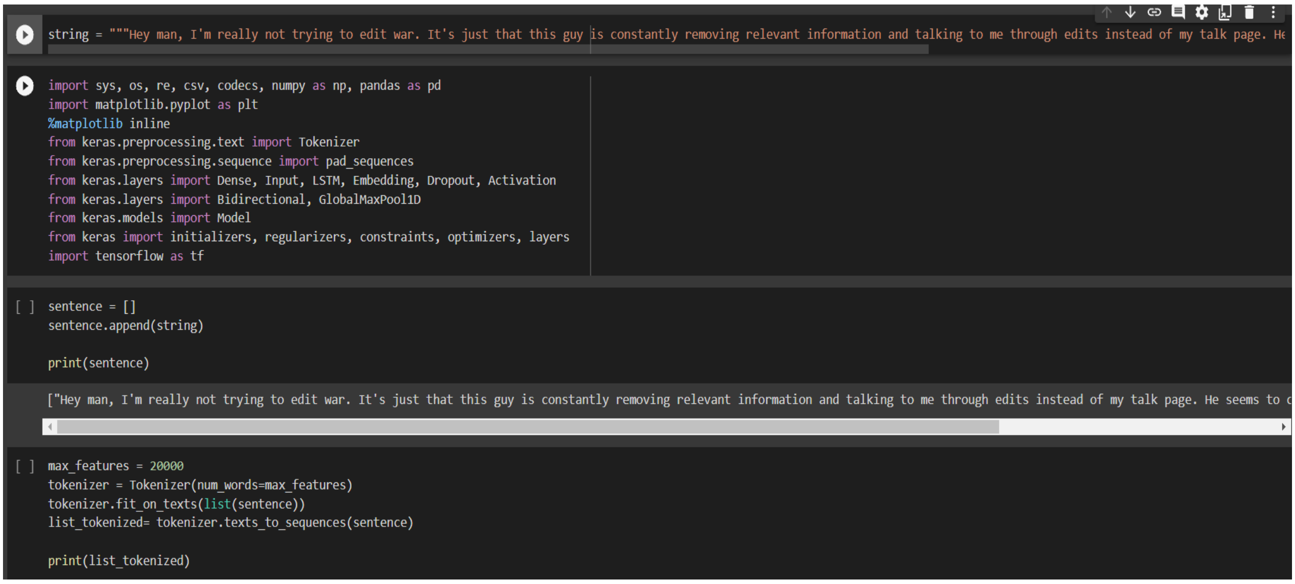
Hardware

- Intel i5 processor
- RAM – 8GB
- Hard disk – 10GB
- Internet Connection

Software

- Python
- Keras Tensorflow
- Pandas
- Numpy
- Google Colab
- Windows

2.5 Experiment and Results for Validation and Verification



```
string = """Hey man, I'm really not trying to edit war. It's just that this guy is constantly removing relevant information and talking to me through edits instead of my talk page. He
```

```
import sys, os, re, csv, codecs, numpy as np, pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
from keras.layers import Dense, Input, LSTM, Embedding, Dropout, Activation
from keras.layers import Bidirectional, GlobalMaxPool1D
from keras.models import Model
from keras import initializers, regularizers, constraints, optimizers, layers
import tensorflow as tf
```

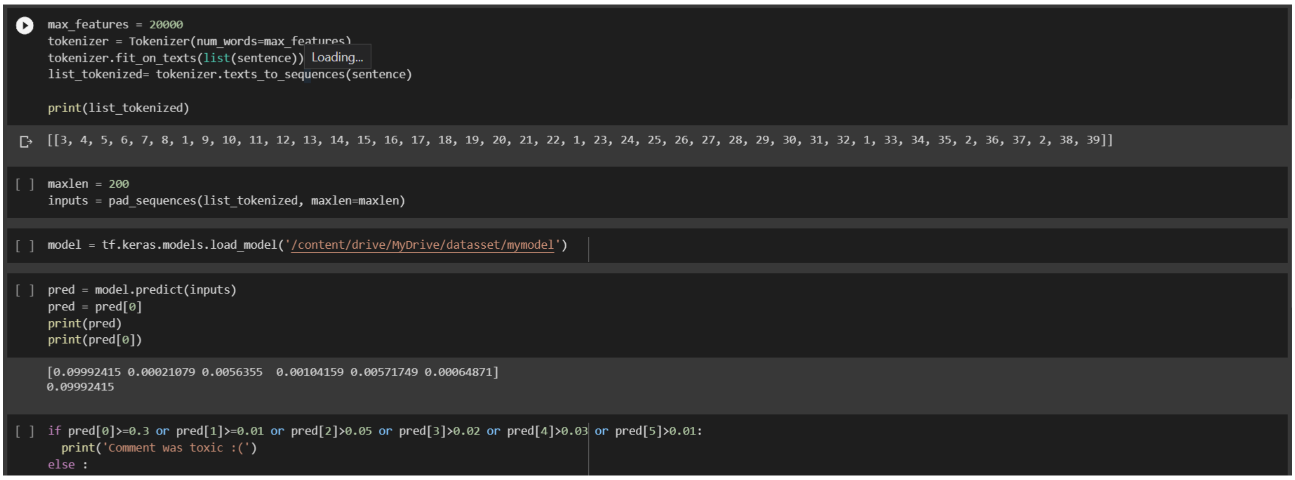
```
sentence = []
sentence.append(string)

print(sentence)
```

```
["Hey man, I'm really not trying to edit war. It's just that this guy is constantly removing relevant information and talking to me through edits instead of my talk page. He seems to c
```

```
max_features = 20000
tokenizer = Tokenizer(num_words=max_features)
tokenizer.fit_on_texts(list(sentence))
list_tokenized= tokenizer.texts_to_sequences(sentence)

print(list_tokenized)
```



```
max_features = 20000
tokenizer = Tokenizer(num_words=max_features)
tokenizer.fit_on_texts(list(sentence)) Loading...
list_tokenized= tokenizer.texts_to_sequences(sentence)

print(list_tokenized)
```

```
[[3, 4, 5, 6, 7, 8, 1, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 1, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 1, 33, 34, 35, 2, 36, 37, 2, 38, 39]]
```

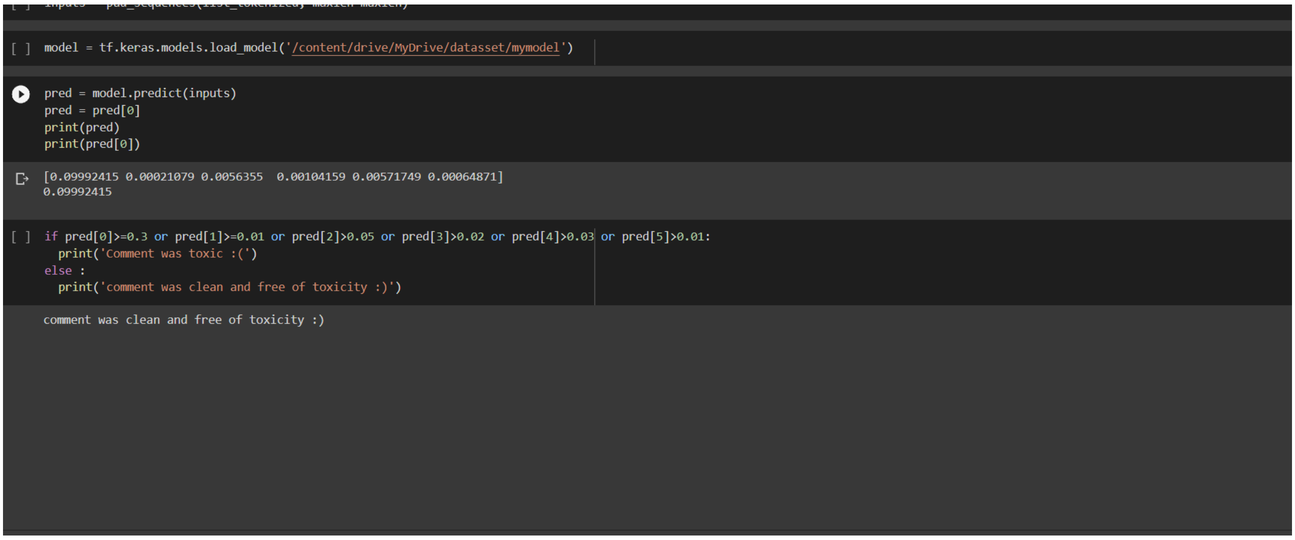
```
maxlen = 200
inputs = pad_sequences(list_tokenized, maxlen=maxlen)
```

```
model = tf.keras.models.load_model('/content/drive/MyDrive/datasset/mymodel')
```

```
pred = model.predict(inputs)
pred = pred[0]
print(pred)
print(pred[0])
```

```
[0.09992415 0.00021079 0.0056355  0.00104159 0.00571749 0.00064871]
0.09992415
```

```
if pred[0]>=0.3 or pred[1]>=0.01 or pred[2]>0.05 or pred[3]>0.02 or pred[4]>0.03 or pred[5]>0.01:
  print('Comment was toxic :(')
else :
```



```
inputs = pad_sequences(list_tokenized, maxlen=maxlen)
```

```
model = tf.keras.models.load_model('/content/drive/MyDrive/datasset/mymodel')
```

```
pred = model.predict(inputs)
pred = pred[0]
print(pred)
print(pred[0])
```

```
[0.09992415 0.00021079 0.0056355  0.00104159 0.00571749 0.00064871]
0.09992415
```

```
if pred[0]>=0.3 or pred[1]>=0.01 or pred[2]>0.05 or pred[3]>0.02 or pred[4]>0.03 or pred[5]>0.01:
  print('Comment was toxic :(')
else :
  print('comment was clean and free of toxicity :)')
```

```
comment was clean and free of toxicity :)
```

2.6 Conclusion and Future Scope

The current situation regarding Internet Toxicity has been critical. Posting comments in online discussions has become an important way to exercise one's right to freedom of expression on the web. These essential rights however are under attack: malicious users hinder otherwise respectful discus-sions with their toxic comments. A toxic comment is defined as a rude, dis-respectful, or unreasonable comment that is likely to make other users leave discussion.

This project will reduce the help in tackling this pressing issue. Within this system, the users can help in eliminating toxicity and build a healthy and free internet community.

2.7 References

- https://keras.io/
- https://www.kaggle.com/
- www.geeksforgeeks.org/