

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL



Department of INFORMATION TECHNOLOGY

IT890 - Professional Practice/ Seminar

Comment Toxicity Detection

Under The Guidance of
Dr. Shrutilipi Bhattacharjee

Submitted By: Nikhil Verma (222IT026)



● Comment Toxicity Detection ●



Introduction

- Comment toxicity refers to the presence of harmful, offensive, or inappropriate language in online comments, which can lead to negative impacts on individuals and communities.
- This project focuses on training a accurate model to classify comment into different levels of toxicity.
- Create API for toxic comment detection.

Motivation

Life in 21st Century

2



Motivation

Negative Effects of Social Media

3

More children 'self-harming' because of cyber-bullying

THE number of children admitted to hospital for self-harm has risen by 30 per cent in a year - with cyber-bullying in blame, according to experts.

Children included in the study from Plymouth included a boy of seven from Devon who tried to hang himself, a 12-year-old girl who tried to stab herself with a knife, and a 13-year-old girl who tried to stab herself with a knife.

It represents a rise on 10 per cent from 2012, when 1,000 children were admitted to hospital for self-harm.

Cyberbullying knows no age limit

Adults too need to learn a thing or two about behaving online, so they don't cause harm to others.



It's a common mistake to think that cyberbullying is only a problem for children. In fact, it can affect anyone who uses the internet. And it's not just about being bullied - it's also about being a bully.

One woman, who asked not to be named, told me that she had been bullied online for years. She said that the bullying had started when she was 15 and continued until she was 25. She said that the bullying had caused her to lose her job and her home.

She said that the bullying had started when she was 15 and continued until she was 25. She said that the bullying had caused her to lose her job and her home.



New cyber-bullying weapon: Mobile phones

Widespread use of handsets make youths an easy target

THE CCGS NEWS

'CYBERBULLYING RUINES LIVES'



WHY SHOULD WE CARE ABOUT CYBERBULLYING? Because it can lead to drugs or drink abuse. Because it can lead to self-harm. Because it can lead to suicide. Because it can lead to a loss of identity. Because it can lead to a loss of control. Because it can lead to a loss of freedom. Because it can lead to a loss of hope. Because it can lead to a loss of love. Because it can lead to a loss of life.

Death by social media

South Africa has a cyberbullying rate of 24% which places it at number four in the world

Cyberbullying has become a global problem. It is a form of bullying that takes place through the use of technology. It can be as simple as sending a mean text message or as serious as posting a false rumor about someone. Cyberbullying can have serious consequences for the victims. It can lead to depression, anxiety, and even suicide.

In South Africa, the cyberbullying rate is 24%, which places it at number four in the world. This is a concerning trend, especially given the high rates of cyberbullying in other countries. It is important that we take steps to prevent cyberbullying and protect our children from its harmful effects.

Missouri woman indicted in case involving MySpace-related suicide

By Linda Simon

A Missouri woman has been indicted in a case involving the suicide of a young woman who was allegedly bullied on MySpace.

The woman, who is 34 years old, is charged with first-degree murder in the death of a 17-year-old girl. The girl was found dead in a wooded area near her home in St. Louis.

The girl had been bullied on MySpace for several months before her death. The woman, who is the girl's mother, is accused of encouraging the girl to commit suicide.

The case has drawn widespread attention, as it highlights the dangers of cyberbullying and the potential for online harassment to lead to real-world violence.

Online abuse becoming part of kids' life

4 In 10 Parents Unable To Help; Schools Urged To Teach How To Tackle Cyberbullying

By [Name]

Online abuse is becoming a part of kids' life, and parents are often unable to help. A new study shows that 4 in 10 parents are unable to help their children deal with online abuse. The study also found that schools are often unable to teach children how to deal with online abuse.

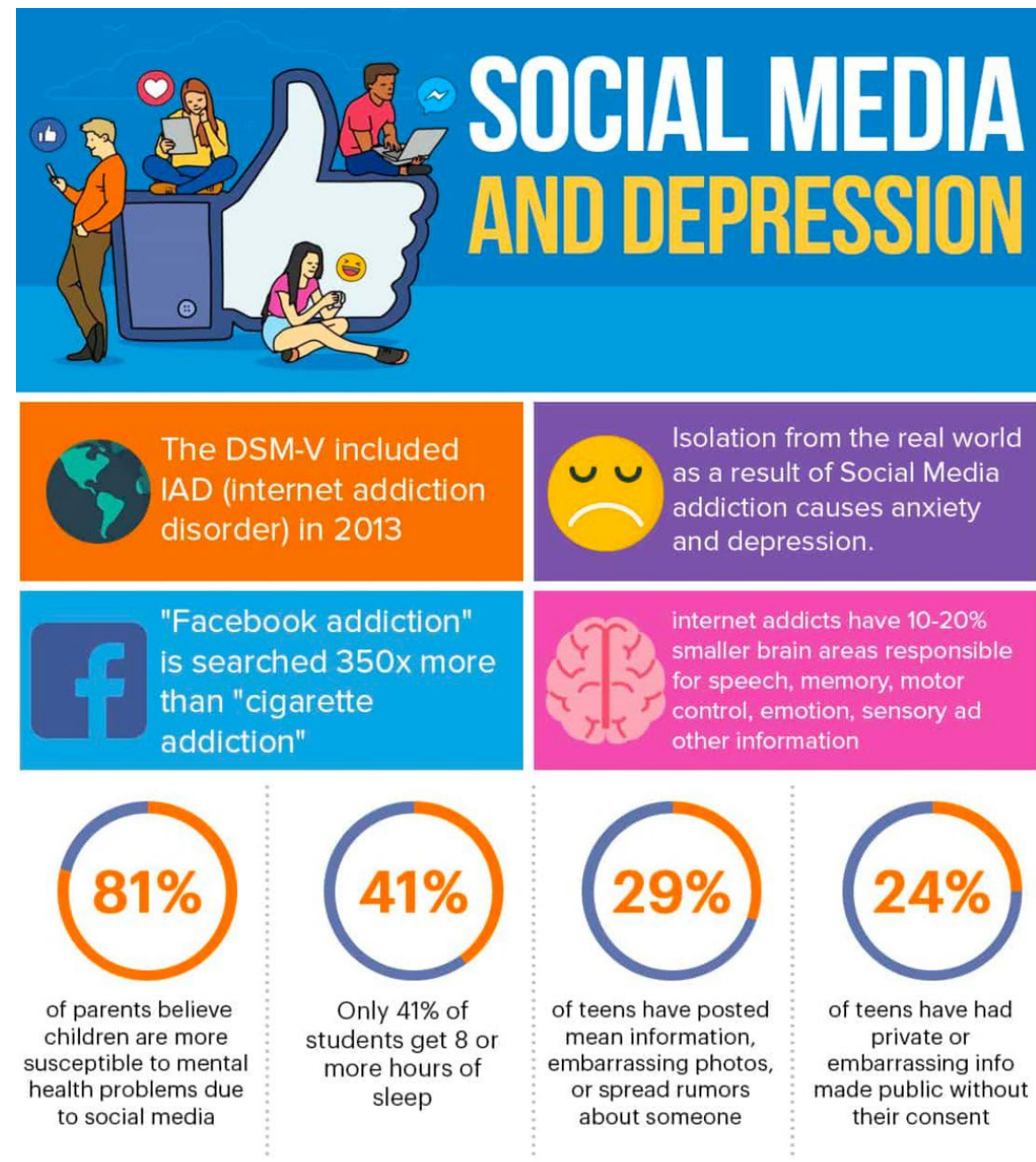
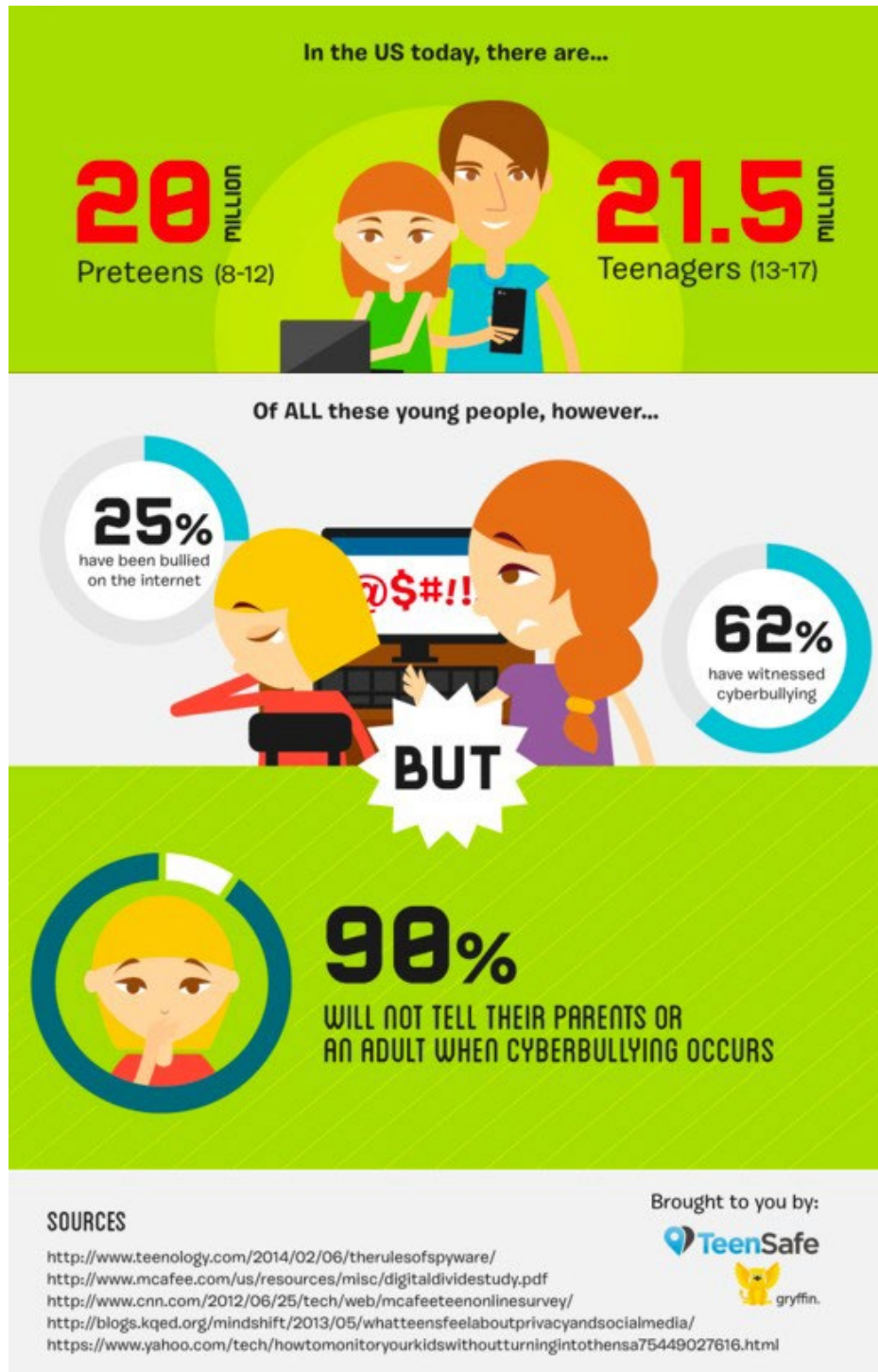
The study, which was conducted by a team of researchers from the University of Michigan, found that 4 in 10 parents are unable to help their children deal with online abuse. The study also found that schools are often unable to teach children how to deal with online abuse.

The researchers say that this is a concerning trend, as online abuse can have serious consequences for children. It can lead to depression, anxiety, and even suicide. They urge parents and schools to take steps to prevent online abuse and protect children from its harmful effects.

Motivation

Surveys

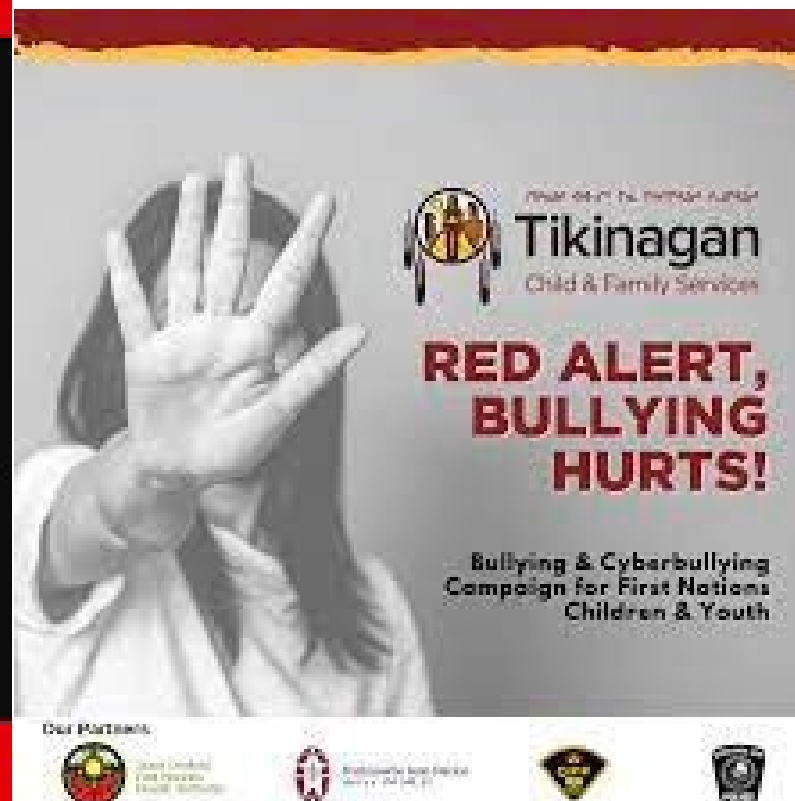
4



Motivation

Preventive Measures

5



Literature Survey

Title	Year & publication	Method	Data set	Result
Convolutional neural networks for toxic comment classification	10th hellenic conference on artificial intelligence. 2018.	CNN for classifying into toxic and non toxic category	Wikipedia talk page	91.2 %
Machine learning methods for toxic comment classification: a systematic review	Acta Universitatis Sapientiae, Informatica, vol.12, no.2, 2020	Study of various machine learning and deep learning model	Twitter Dataset	
A Supervised Multi-class Multi-label Word Embeddings Approach for Toxic Comment Classification.	International conference on knowledge and information retrieval, 2019	Study of various machine learning and deep learning model	Wikipedia talk page	92 %
DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter	arXiv, 2019	Distilled version of BERT	Wikipedia talk page	40% fewer parameters than BERT and is 60% faster than BERT.

Outcome of literature survey

- CNN is used for binary classification into toxic and non toxic comments is performing good and it is computationally less expensive.
- Different technique for text embedding.
- Text representation technique that capture the meaning of words in context of surrounding words in sentence.

Problem Statement

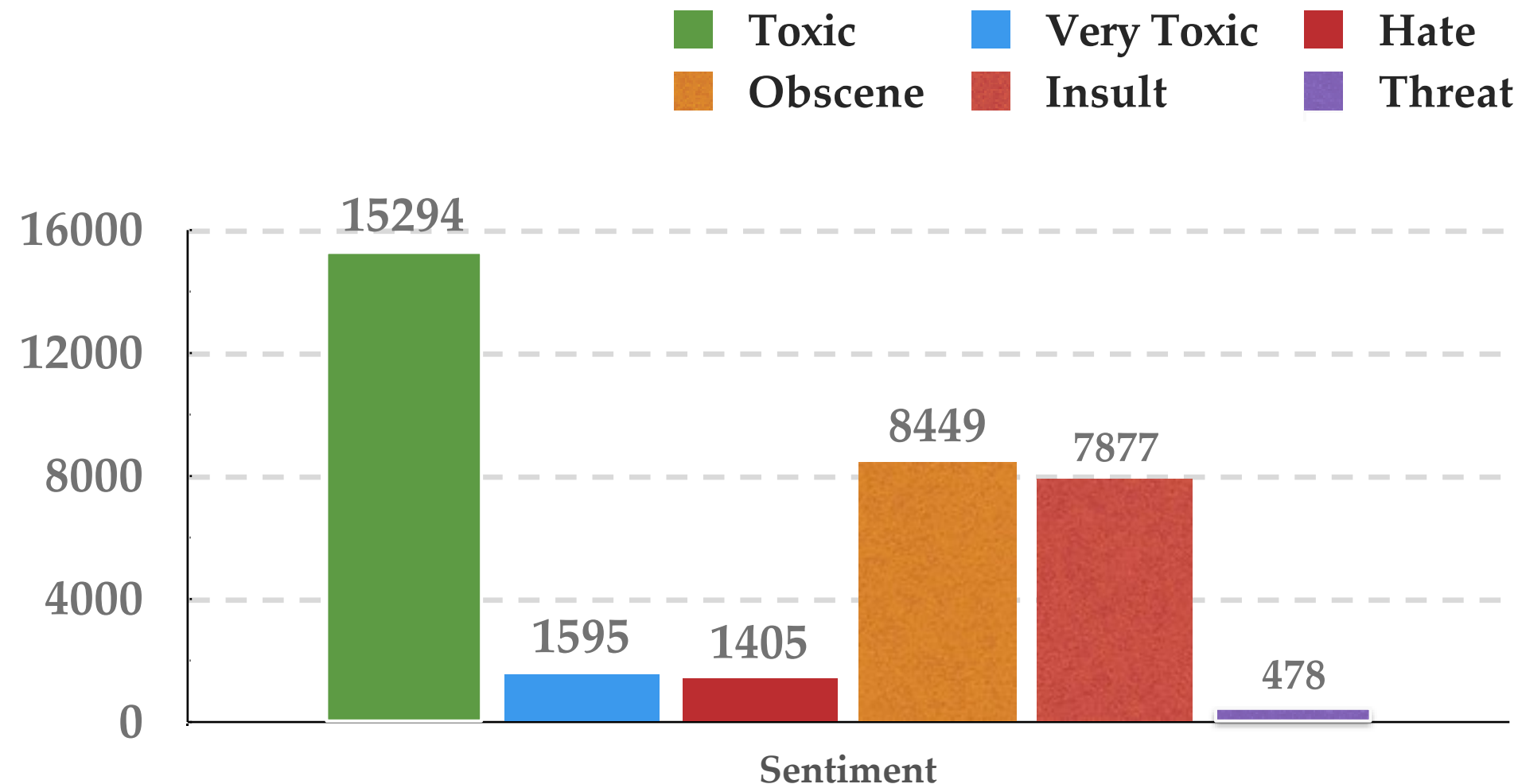
- To train a comment toxicity detection model.
- Create a Web API that can perform classification based on the saved trained model.

Methodology

- Data Analysis
- Data Preprocessing
- Data Vectorization
- LSTM, GRU, Logistic Regression model implementation
- Model Evaluation
- API implementation

Exploratory Data Analysis (EDA)

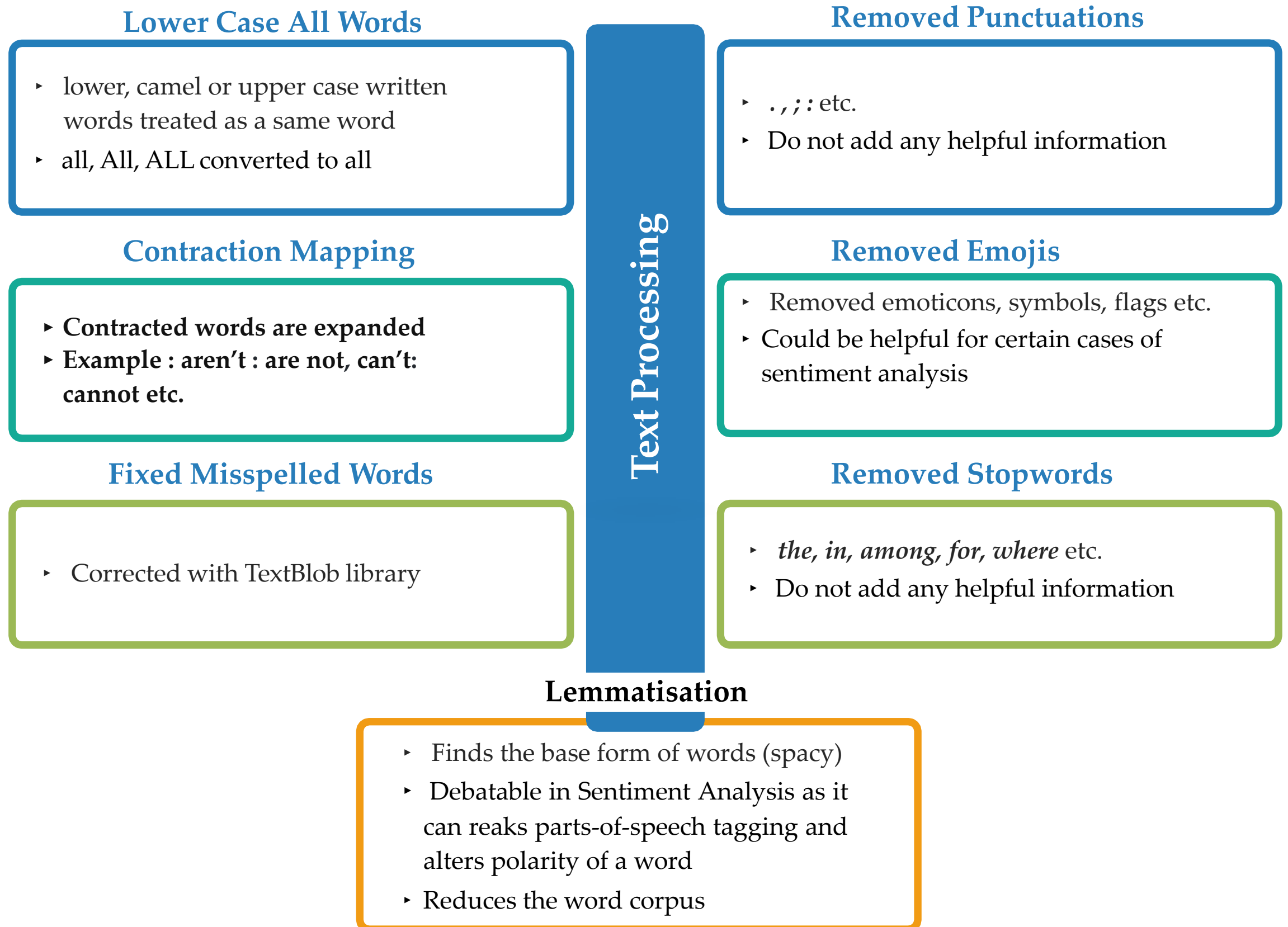
10



- Training Data: Total: 159571 entries, Neutral: 124473 (none of the toxicity labels assigned)
- Huge unbalanced dataset, no null entries and empty strings present
- Multi-Label Classification Problem, target labels are not mutually exclusive, i.e. More than one right answer

Data Preprocessing

11



Data Preprocessing

12

Embedding Vectors

- Pre-trained word embedding vectors [Glove.6B](#) is used

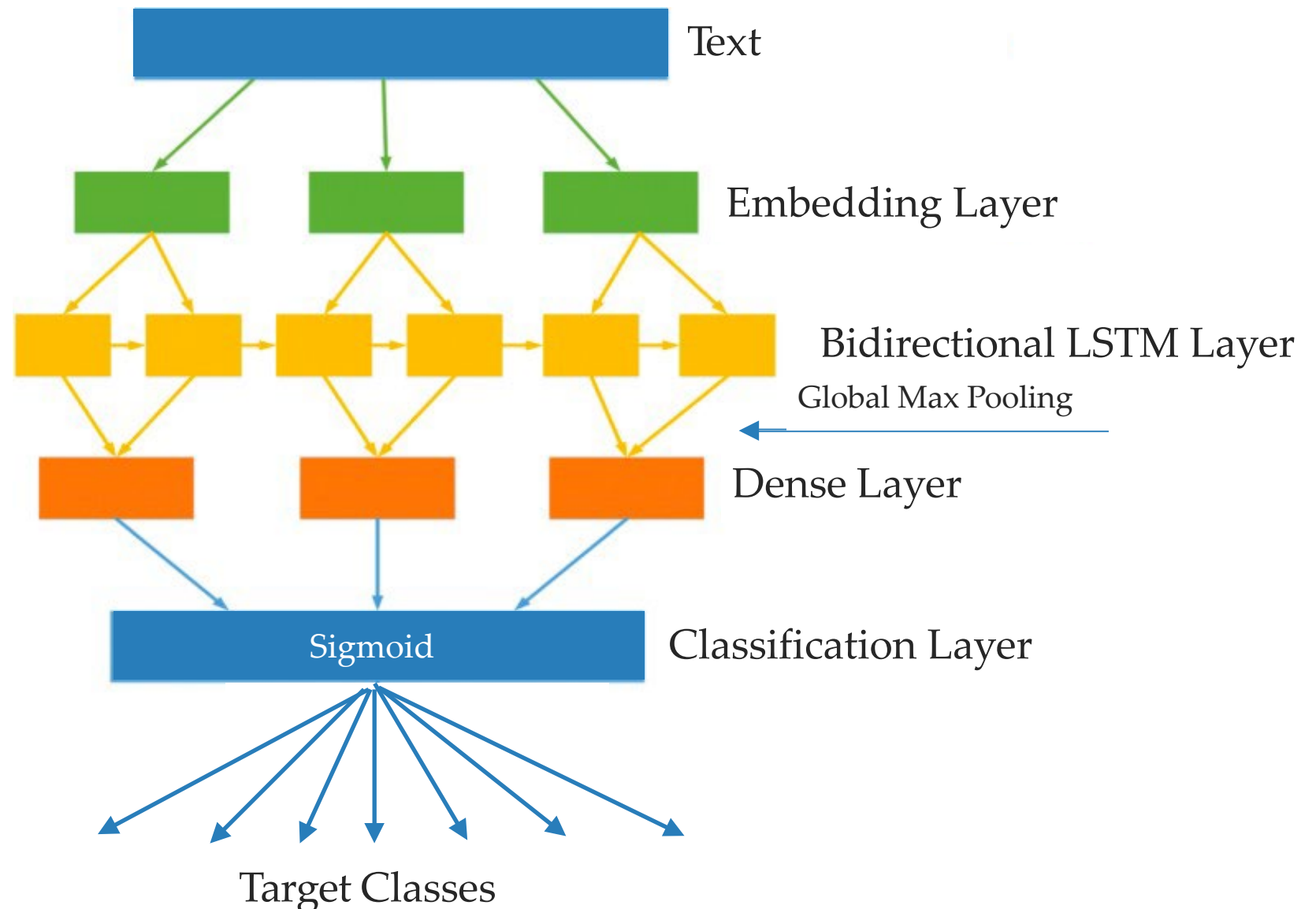
Tokenizer

- Tokenizer (from keras) is used to tokenize the text with max vocab size of 20000
- Each word is assigned to the corresponding feature vector from the Glove.6B

Deep Learning Model (Bi-directional LSTM)

13

- Recurrent Neural Network (RNN) is used as it was primarily built to tackle NLP problems (sequential data, sentences are sequence of words)
- I have used Bidirectional Long Short-Term Memory (LSTM) RNN model.
- LSTM can remember longer sequences than regular RNN
- Making it bidirectional, helps in a way that it can see at a given sequence both previous and next sequences in the text/ sentence.



Experimental Setup

- Language : PYTHON
- IDE : Anaconda, PyCharm, Kaggle Notebook
- Model Building : tensorflow, keras
- Web API : GRADIO

Results

Logistic Regression Model Evaluation using 5-fold cross validation

16

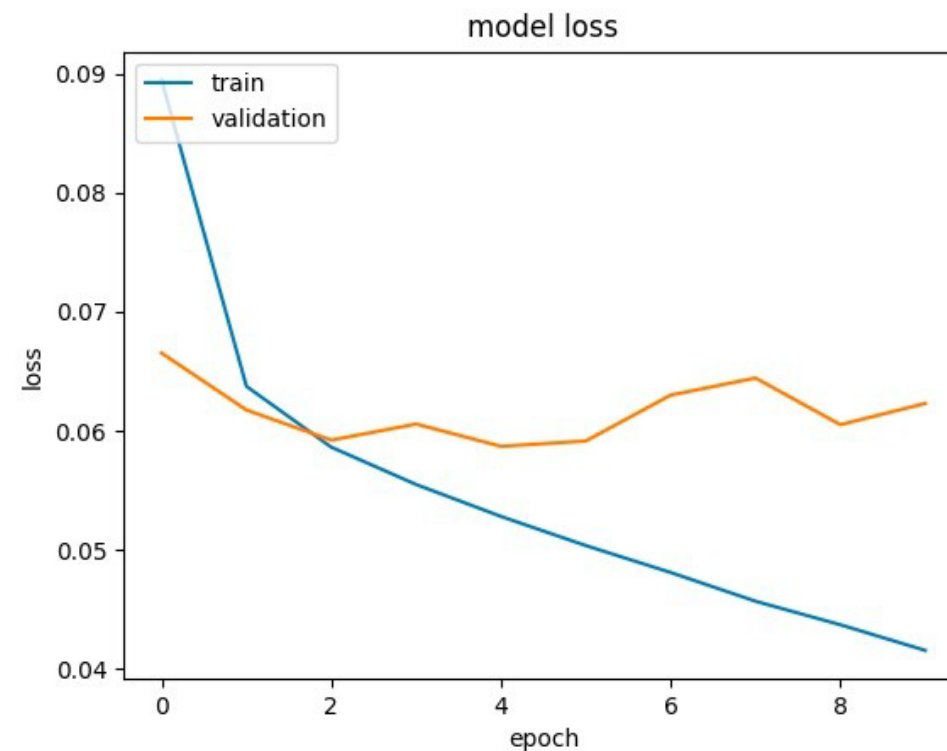
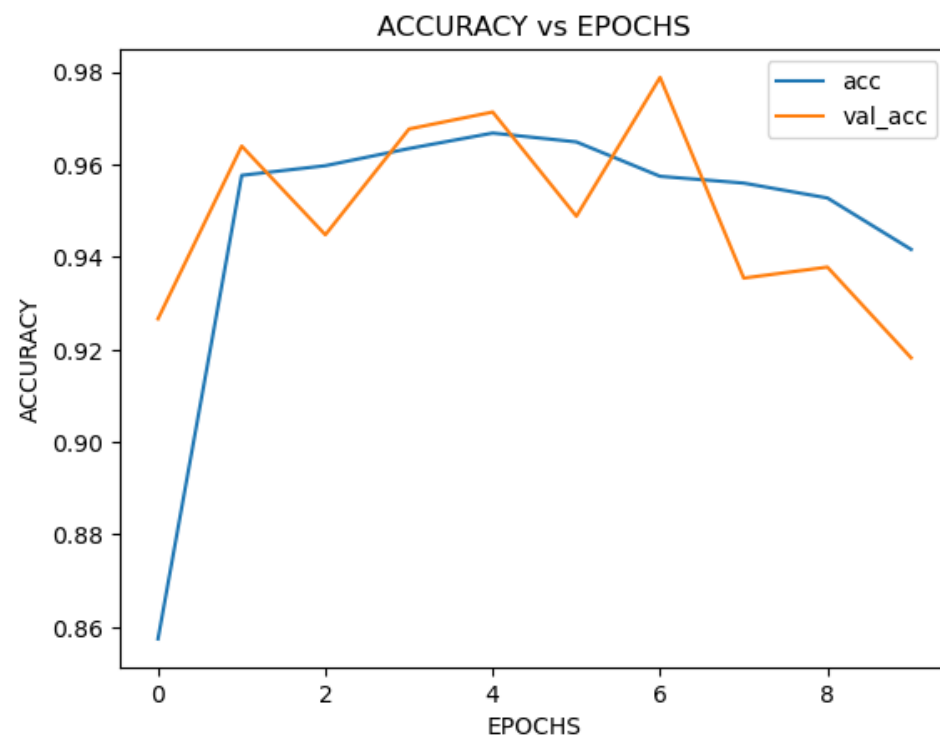
- The logistic Regression model is trained and evaluated for each class separately as logistic regression model can only perform binary classification. the average accuracy of each class is 98.59%.
- Performance metrics for Logistic Regression with 5-fold cross validation

Metrics	Score
Area Under Curve (AUC)	0.985991
F1 Score	0.6701
Accuracy	98.59 %

Bi-directional GRU Model Evaluation

17

- 20% of training dataset is kept for validation and rest used for training
- Accuracies on both training and validation sets and loss in each epoch (used early stopping monitoring validation loss)

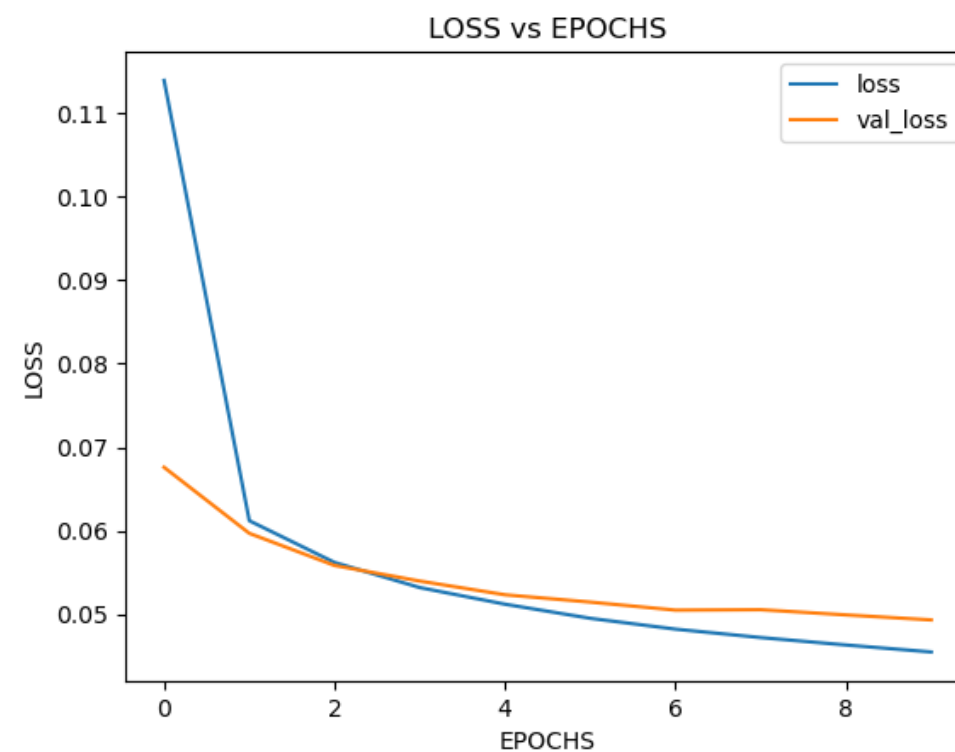
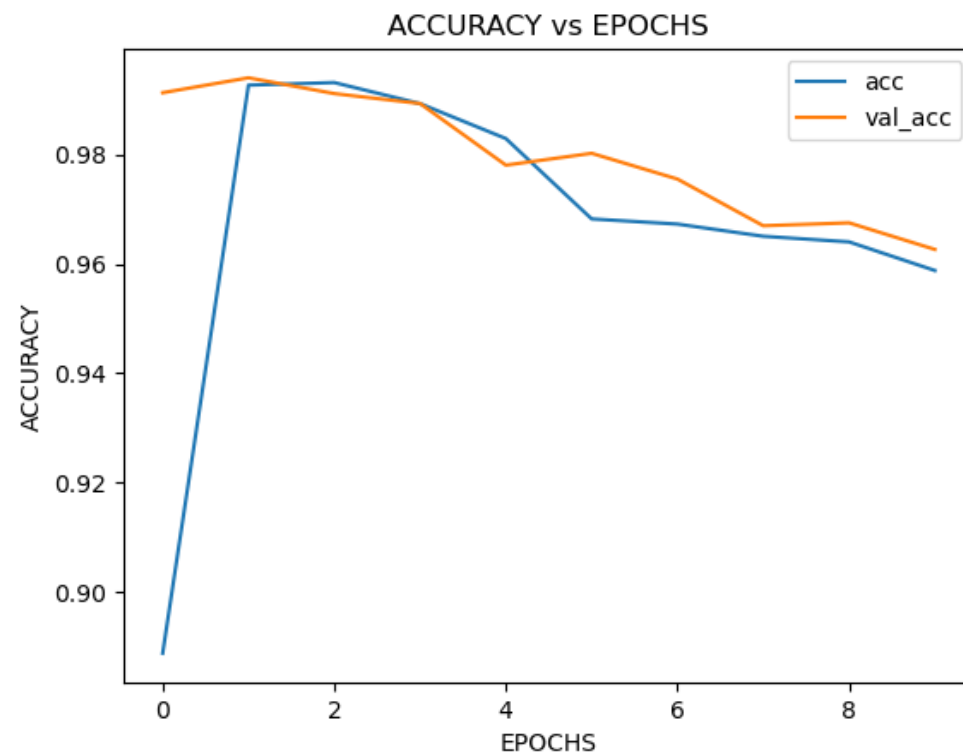


Metrics	Score
Validation Loss	0.0497
Validation Accuracy	91.82%

Bi-directional LSTM Model Evaluation

18

- 20% of training dataset is kept for validation and rest used for training
- Accuracies on both training and validation sets and loss in each epoch (used early stopping monitoring validation loss)



Metrics	Score
Validation Loss	0.0493
Validation Accuracy	96.26%

DEMO

Conclusion & Future Work

- Bi-directional LSTM model is the best performing model with 96 % Accuracy.
- API is correctly classifying the text into different level of toxicity.

Additionally, the followings are some suggested studies to be considered as future work in this area:

- Create Comment Toxicity detection for Multilingual Text.
- using Other DNN techniques (CNN)) because some recently published papers s have shown that CNN proves to have a very high performance for various NLP tasks.
- Using other text embedding and performing Hyper parameter tuning to improve performance of machine learning model.

1. Georgakopoulos, Spiros V., et al. "Convolutional neural networks for toxic comment classification." Proceedings of the 10th hellenic conference on artificial intelligence. 2018.
2. Androćec, Darko. "Machine learning methods for toxic comment classification: a systematic review" *Acta Universitatis Sapientiae, Informatica*, vol.12, no.2, 2020, pp.205-216.
<https://doi.org/10.2478/ausi-2020-0012>.
3. Carta, Salvatore, et al. "A Supervised Multi-class Multi-label Word Embeddings Approach for Toxic Comment Classification." *KDIR*. 2019.
4. Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108* (2019).

Thank You