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

Motivation

Negative Effects of Social Media

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

Cyberbullying knows no age limit

Missouri woman indicted in case involving MySpace-related suicide

CYBER BULLYING AFFECTS REAL LIVES!

New cyber-bullying weapon: Mobile phones

Widespread use of handsets make youths an easy target

THE CCGS NEWS

'CYBERBULLYING RUINES LIFES'

Death by social media

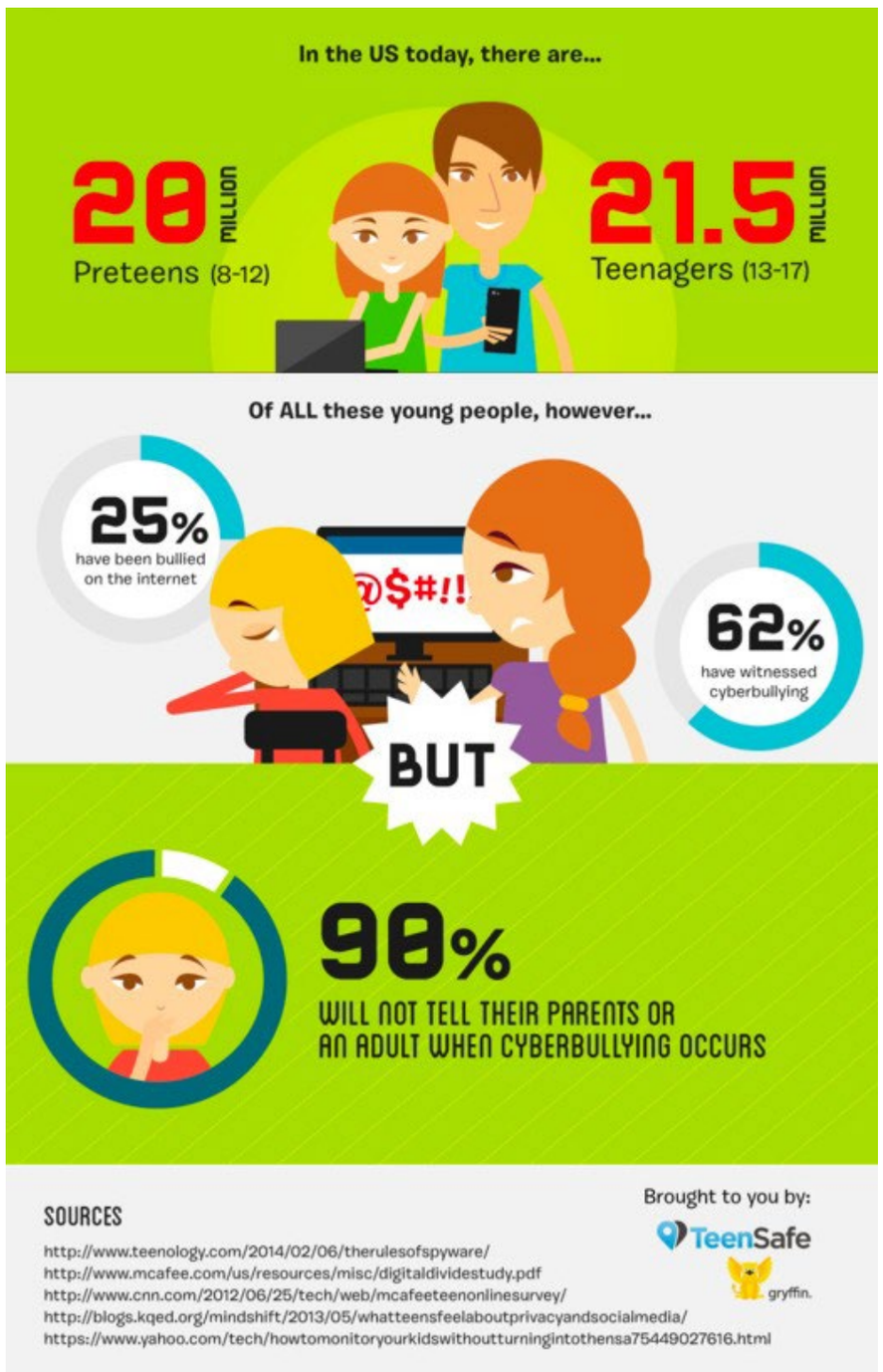
TIMES TRENDS

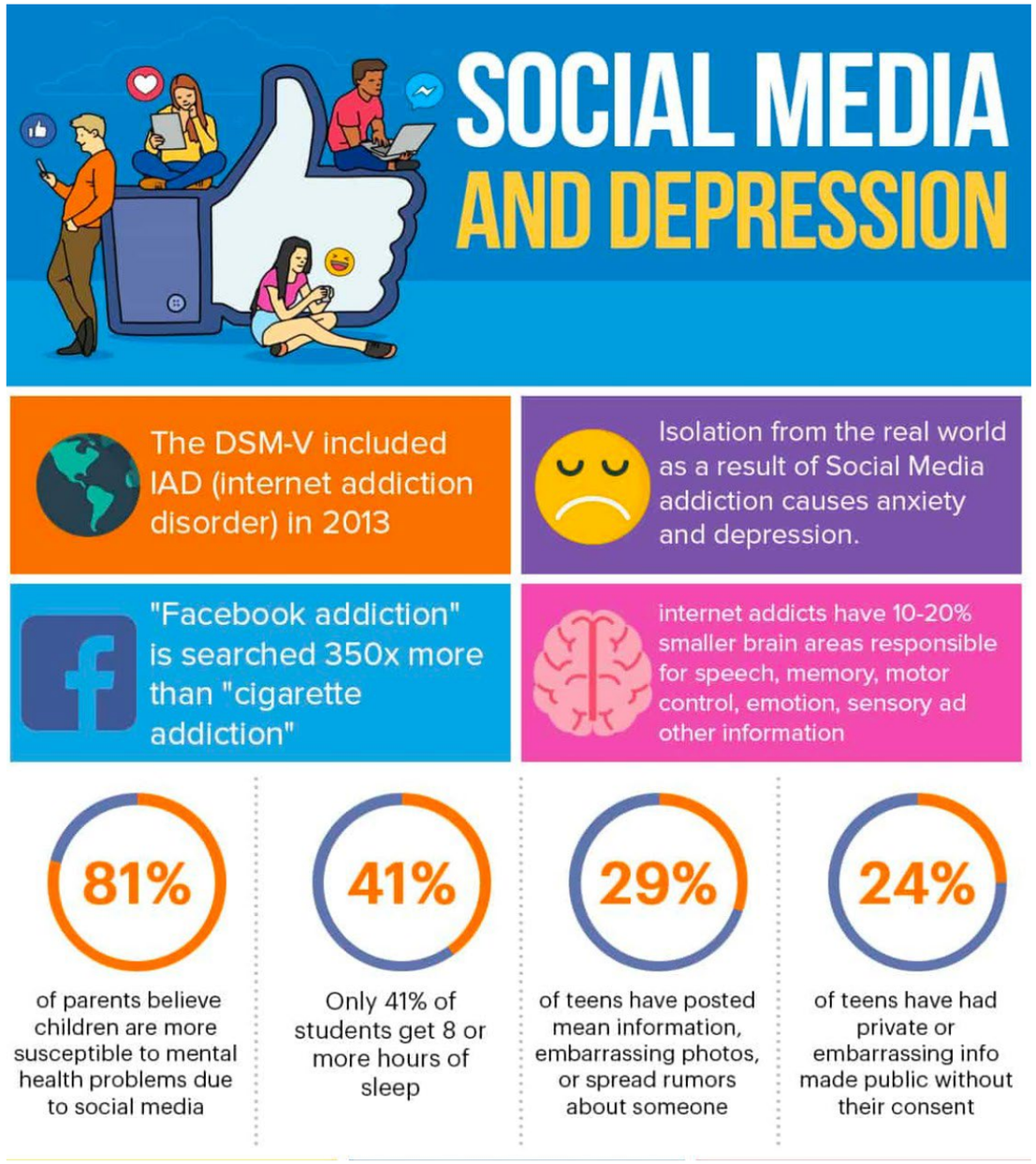
Online abuse becoming part of kids' life

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

Motivation

Surveys





Motivation

Preventive Measures

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.

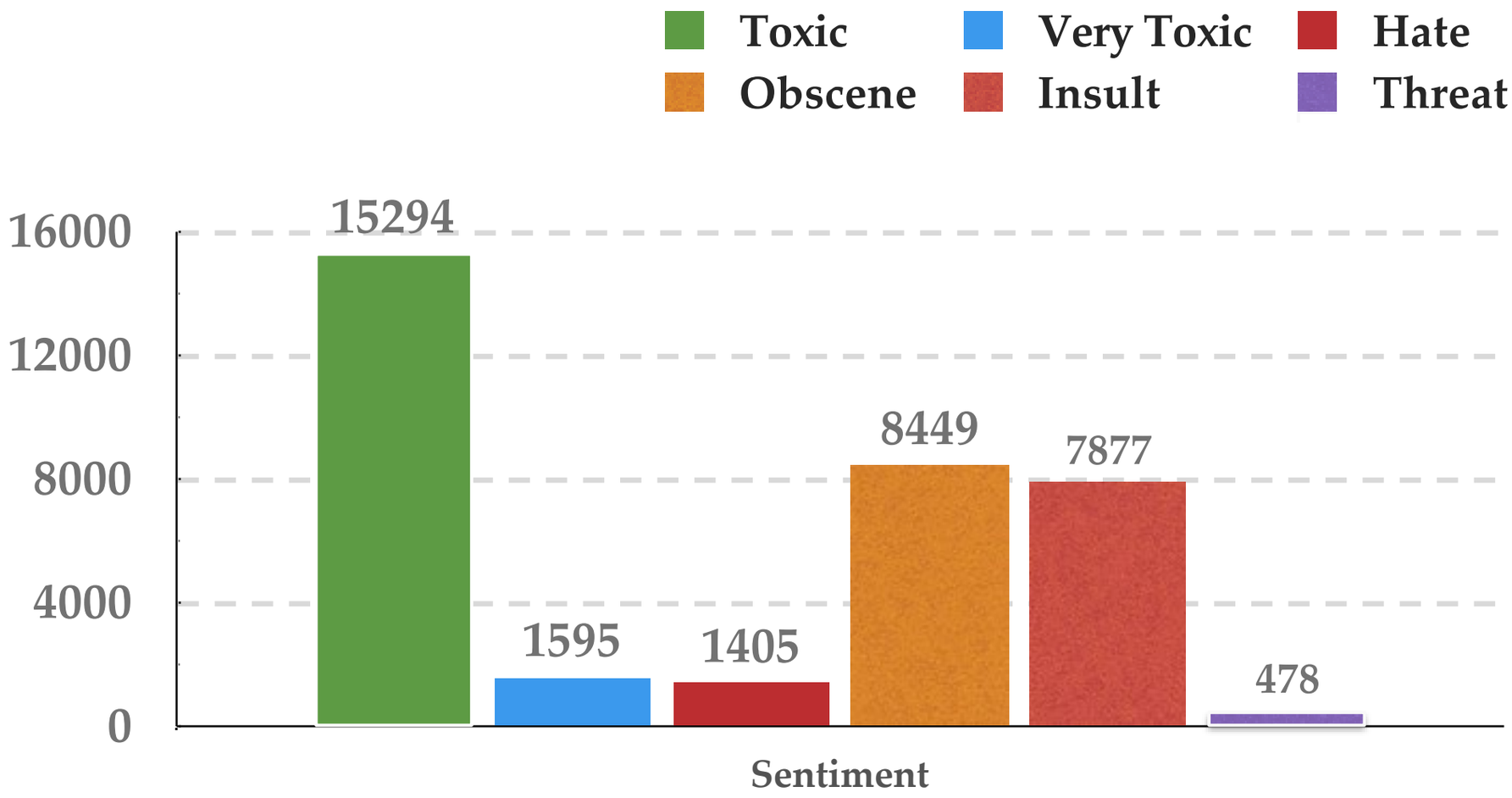
Problem Statement

- 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)



- 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

Lower Case All Words

- lower, camel or upper case written words treated as a same word
- all, All, ALL converted to all

Removed Punctuations

- . , ; : etc.
- Do not add any helpful information

Contraction Mapping

- **Contracted words are expanded**
- **Example : aren't : are not, can't: cannot etc.**

Removed Emojis

- Removed emoticons, symbols, flags etc.
- Could be helpful for certain cases of sentiment analysis

Text Processing

Fixed Misspelled Words

- Corrected with TextBlob library

Removed Stopwords

- *the, in, among, for, where* etc.
- Do not add any helpful information

Lemmatisation

- Finds the base form of words (spacy)
- Debatable in Sentiment Analysis as it can reaks parts-of-speech tagging and alters polarity of a word
- Reduces the word corpus

Data Preprocessing

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

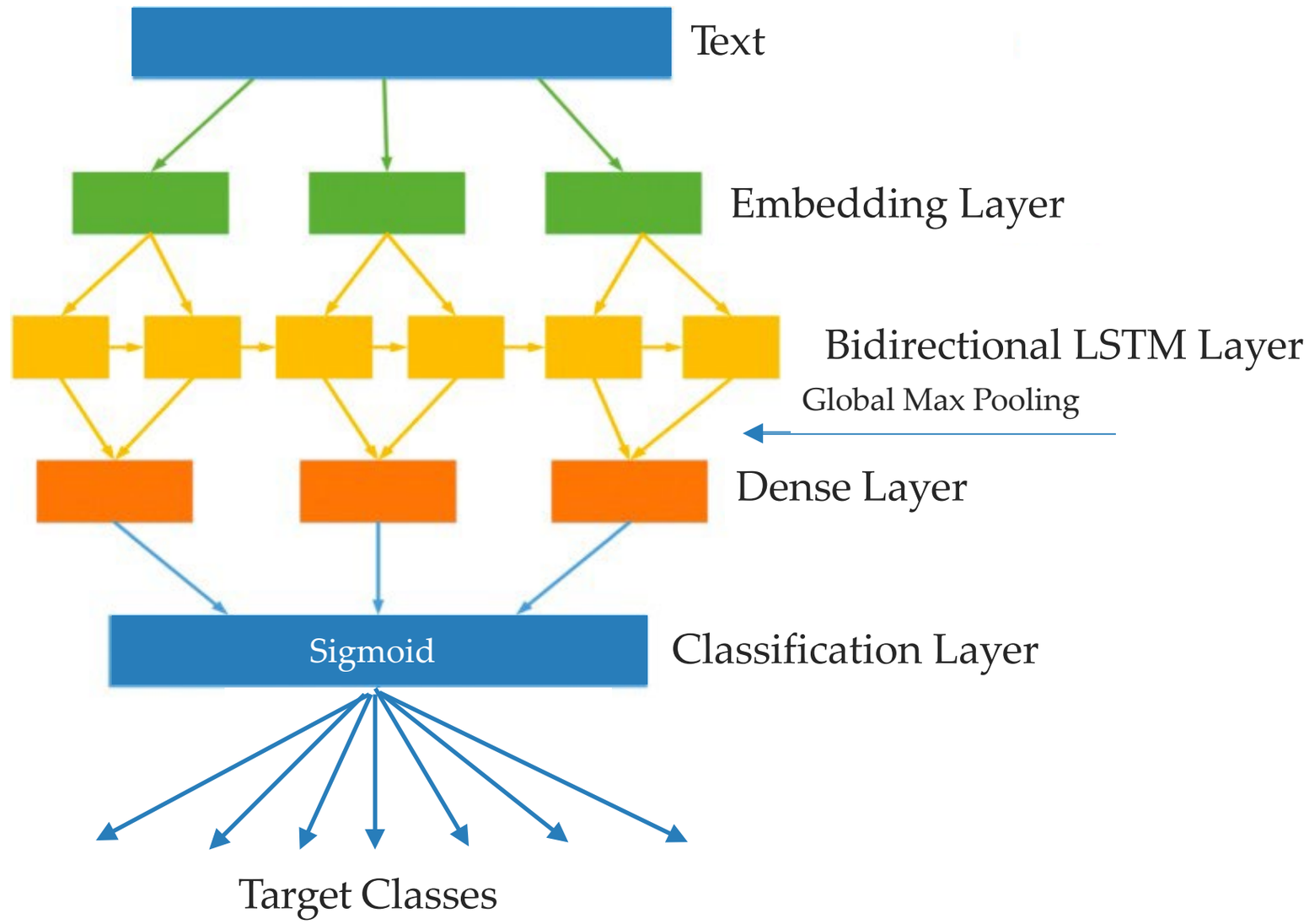
Deep Learning Model (Bi-directional LSTM)

- 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

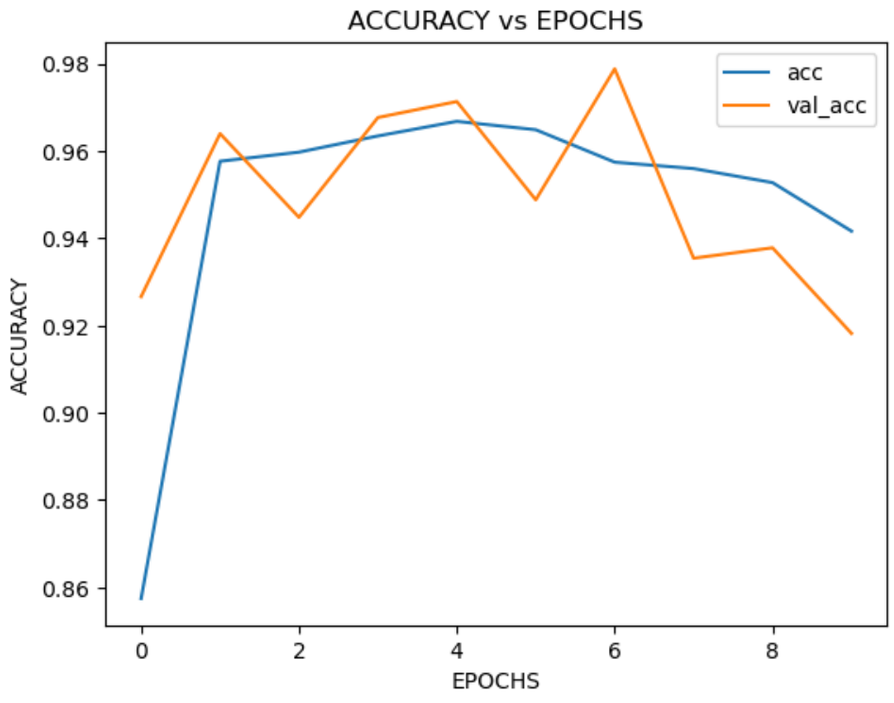
- The logistic Regression model is trained and evaluated for each class separately as logistic regression model can only perform binary classification. the average accuracy of each class is 98.59%.

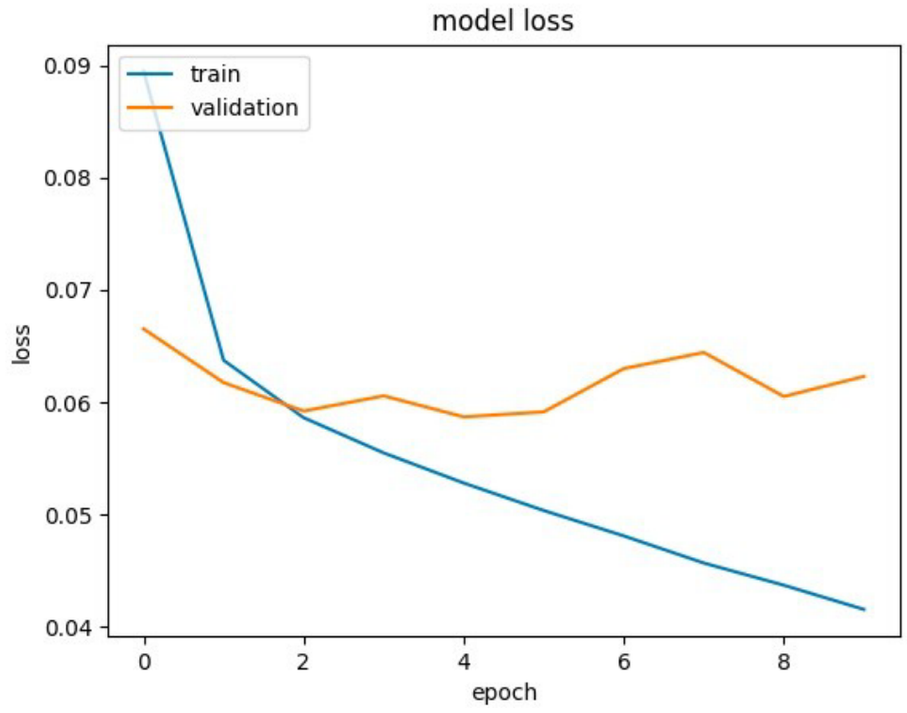
- Performance metrices for Logistic Regression with 5-fold cross validation

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

Bi-directional GRU Model Evaluation

- 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)

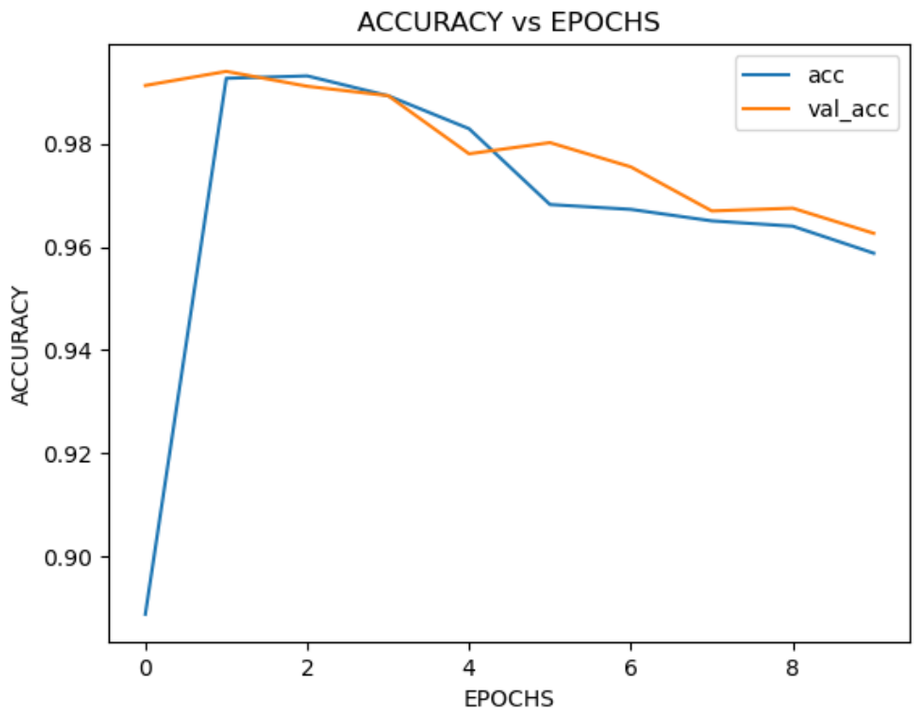


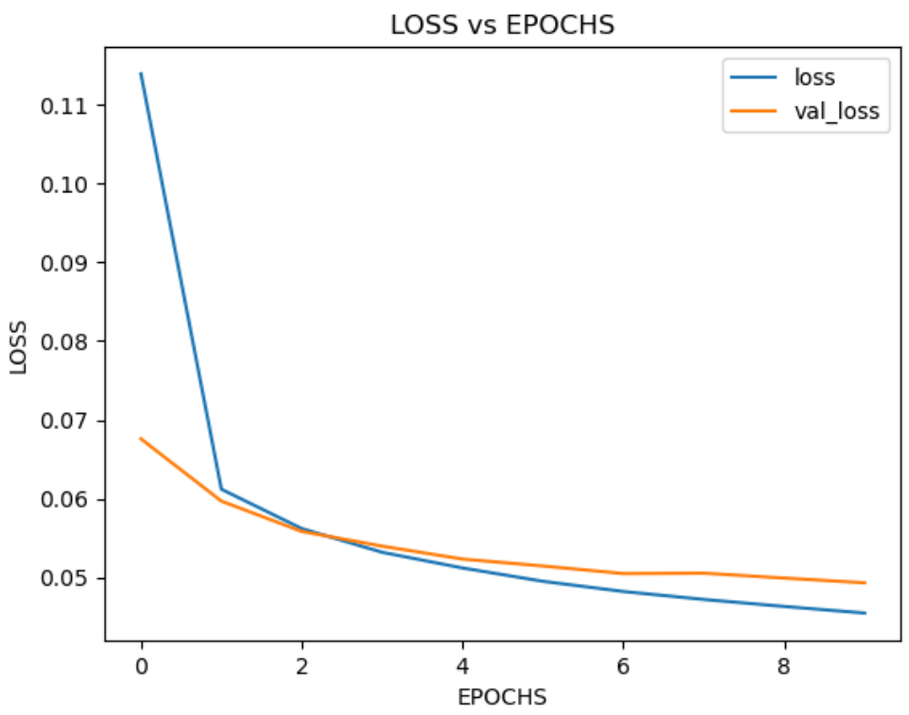


Metrics	Score
Validation Loss	0.0497
Validation Accuracy	91.82%

Bi-directional LSTM Model Evaluation

- 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.

References

1. Georgakopoulos, Spiros V., et al. "Convolutional neural networks for toxic comment classification." Proceedings of the 10th hellenic conference on artificial intelligence. 2018.
2. Andročec, 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