Toxic Comment Classification System using Deep Learning

DIVYABHARATHI L
111722102031 CSE A
09.11.2023

INTRODUCTION

- In recent years, the internet has become a central platform for communication and information sharing.

- However, the rise of toxic comments, including hate speech, abuse, and harassment, poses a significant challenge in maintaining respectful and safe online conversations.

- The Toxic Comment Classification System using Deep Learning project aims to develop an intelligent system capable of automatically identifying and categorizing toxic comments in online discussions.

INTRODUCTION

- With the proliferation of online communication platforms, addressing toxic or harmful content is crucial for maintaining a healthy online environment.

LITERATURE REVIEW

TITLE OF THE PAPER	JOURNAL	AUTHORS	METHODOLOGY	LIMITATIONS
Toxic Comment Classification	IEEE journal	Anton Liu Yue Zhuang Grace Wu	Uses CNN and LSTM in architecture Focuses on multi-labelling classification	Errors due to mislabeling of the training data. The model accuracy was only 50% in certain cases.
Social Media Toxicity Classification Using Deep Learning	Journal of Emerging Technologies and Innovative Research (JETIR)	Hong Fan Wu Du Abdelghani Dahou Ahmed A. Ewees Dalia Yousri	Uses three BERT models to verify its performance.	High Computational Cost Large Model Size

LITERATURE REVIEW

TITLE OF THE PAPER	JOURNAL	AUTHORS	METHODOLOGY	LIMITATIONS
TOXIC COMMENT CLASSIFICATION USING CONVOLUTIONAL AND RECURRENT NEURAL NETWORKS	International Research Journal of Modernization in Engineering Technology and Science	Victor Blanes Marti	The Paper briefs performance comparison between LSTM, CNN and GRU gates in this NLP problem.	Data Quality and Bias Overfitting
Deep Learning Models and Word Embeddings for Toxicity Detection	International Journal of Engineering Applied Sciences and Technology	Danilo Dessì, Diego Reforgiato Recupero and Harald Sack	It analyses deep learning models based on (CNN), and (LSTM)	Data Bias and Generalization Limited Context Understanding

LITERATURE REVIEW

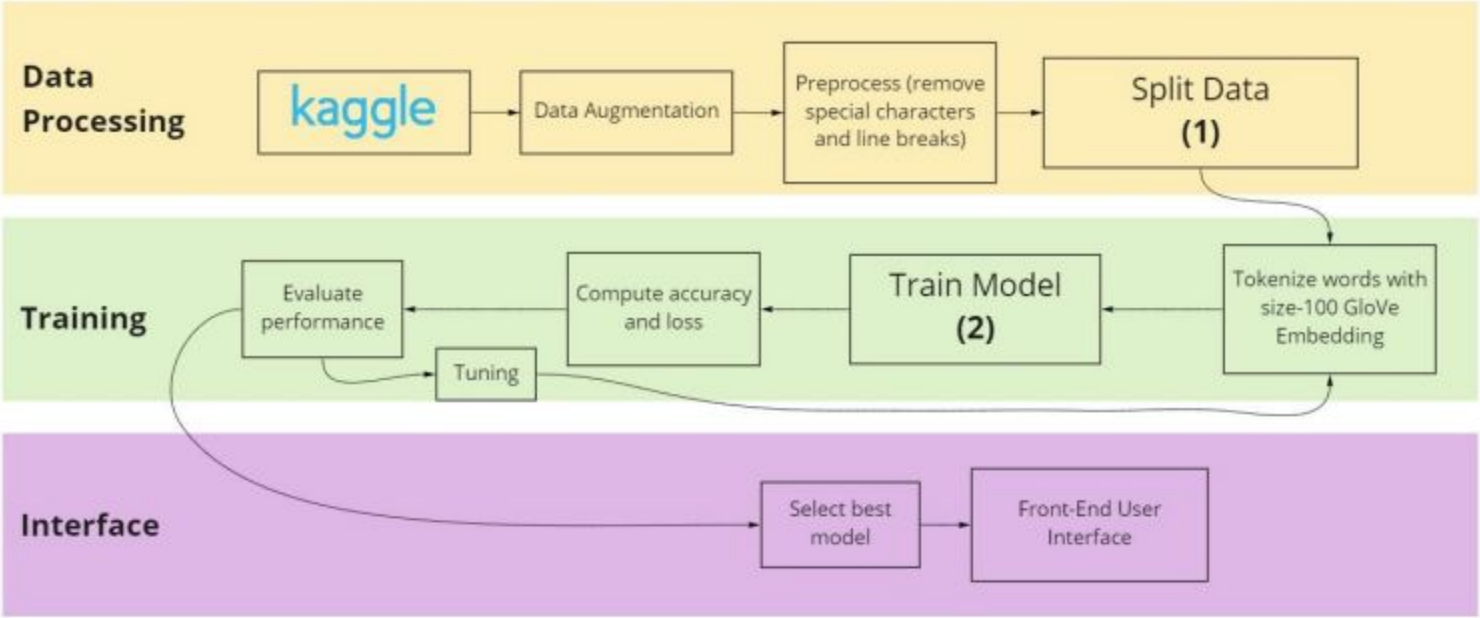
TITLE OF THE PAPER	JOURNAL	AUTHORS	METHODOLOGY	LIMITATIONS
Using a BERT-based Ensemble Network for Abusive Language Detection	27th International Conference on International Conference on Machine Learning	Noah Ballinger	The approach presented in this paper is able to outperform the standard BERT model, and HateBERT, a re-trained variation of the BERT model.	Data Scarcity Large Model Size High Computational Cost

METHODOLOGY

The project is divided into 3 parts :

- ❖ Data processing
- ❖ Training
- ❖ Interface

METHODOLOGY



METHODOLOGY

Data Source :

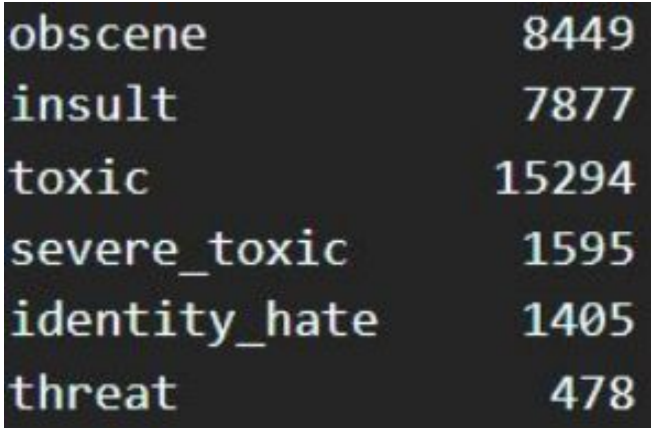
- Data from the “Toxic Comment Classification Challenge” on Kaggle.
- Comments are labelled in 6 categories: toxic, severe toxic, obscene, threat, insult, and identity hate.

Example : If a comment is labelled as 100100, it is toxic and threat.

METHODOLOGY

Examine data:

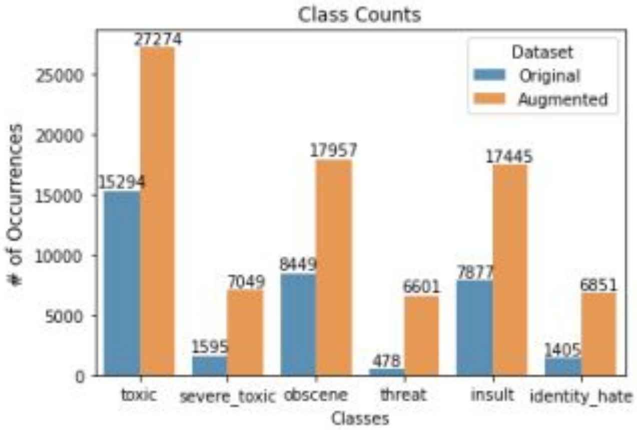
- Examine the data using pandas.
- 150000+ comments are available - both good and bad comments.
- With such low numbers of comments in certain classes, it is easy for the models to overfit and generate a low accuracy for the validation and test set.

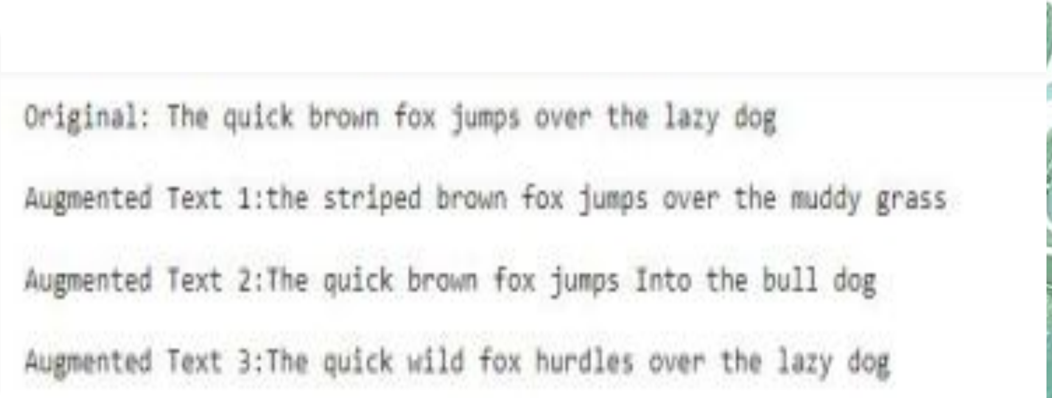


```
obscene           8449
insult            7877
toxic            15294
severe_toxic      1595
identity_hate     1405
threat             478
```

METHODOLOGY

Data Augmentation:
Package nlpaug is used to artificially increase the amount of data in the minority classes by substituting synonyms.





```
Original: The quick brown fox jumps over the lazy dog

Augmented Text 1:the striped brown fox jumps over the muddy grass

Augmented Text 2:The quick brown fox jumps Into the bull dog

Augmented Text 3:The quick wild fox hurdles over the lazy dog
```

METHODOLOGY

Data Cleaning:

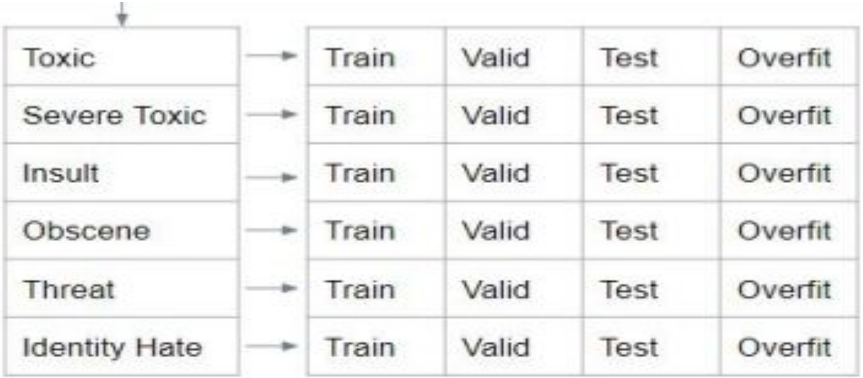
- ❖ The data using regex, matching patterns in the comments and replacing them with more organized counterparts.
- ❖ We removed any spaces, line breaks, contractions, etc.
- ❖ Cleaner data leads to a more efficient model and higher accuracy.

Example: Hello ! Are you there?
hello are you there

METHODOLOGY

Data Processing:

- For prototyping and testing models, n datasets should be created.
- Each with 6 different sub-datasets and each sub-dataset is split into training, validation, testing, and overfit sets.



Toxic	→	Train	Valid	Test	Overfit
Severe Toxic	→	Train	Valid	Test	Overfit
Insult	→	Train	Valid	Test	Overfit
Obscene	→	Train	Valid	Test	Overfit
Threat	→	Train	Valid	Test	Overfit
Identity Hate	→	Train	Valid	Test	Overfit

ARCHITECTURE

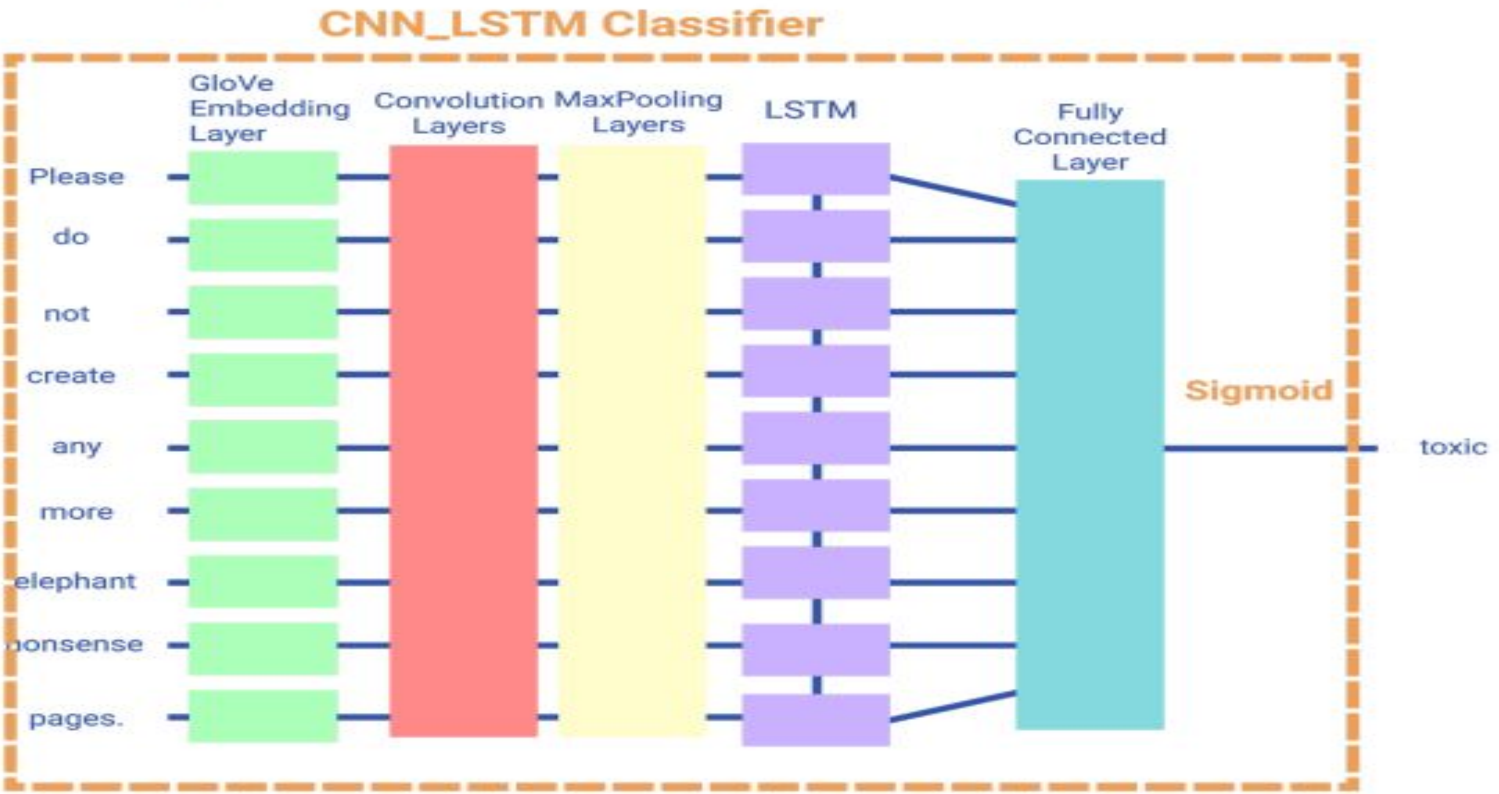
The model used is **CNN_LSTM Binary Classification Model.**

- The input data will be first converted into word vectors with GloVe Embedding.
- Bad comments share similar sets of vocabularies.
- Thus two convolution layers were used to identify word patterns in sentences regardless of their position.
- Each convolution layer has one maxpool layer that gives one output feature for each sentence from each kernel.

ARCHITECTURE

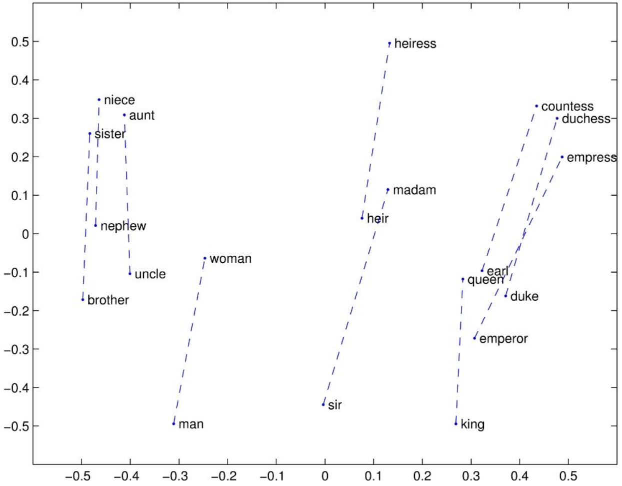
- Then the outputs were concatenated to a vector and were fed to a LSTM (Long Short Term Memory network).
- The output of the LSTM was decoded using a fully connected layer.

ARCHITECTURE



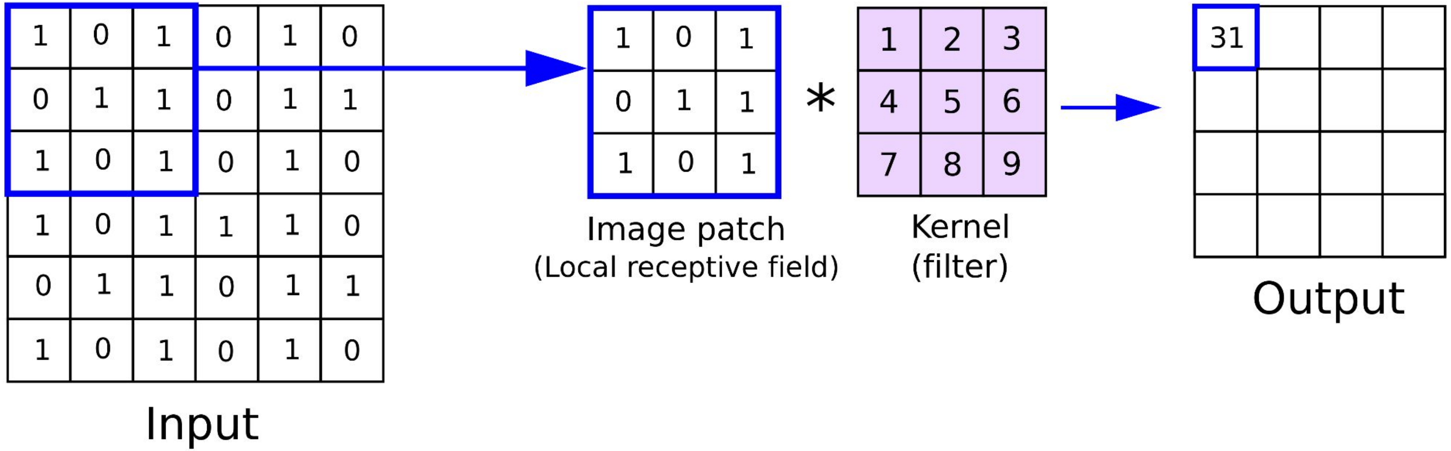
ARCHITECTURE

- GloVe stands for global vectors is an unsupervised learning algorithm for obtaining vector representations for words.
- It map words or phrases from vocabulary to a corresponding vector of real numbers which used to find word predictions, word similarities/semantics.



ARCHITECTURE

- A convolutional layer is the main building block of a CNN.
- It contains a set of filters (or kernels), parameters of which are to be learned throughout the training.



1	0	1	0	1	0
0	1	1	0	1	1
1	0	1	0	1	0
1	0	1	1	1	0
0	1	1	0	1	1
1	0	1	0	1	0

Input

1	0	1
0	1	1
1	0	1

Image patch
(Local receptive field)

*

1	2	3
4	5	6
7	8	9

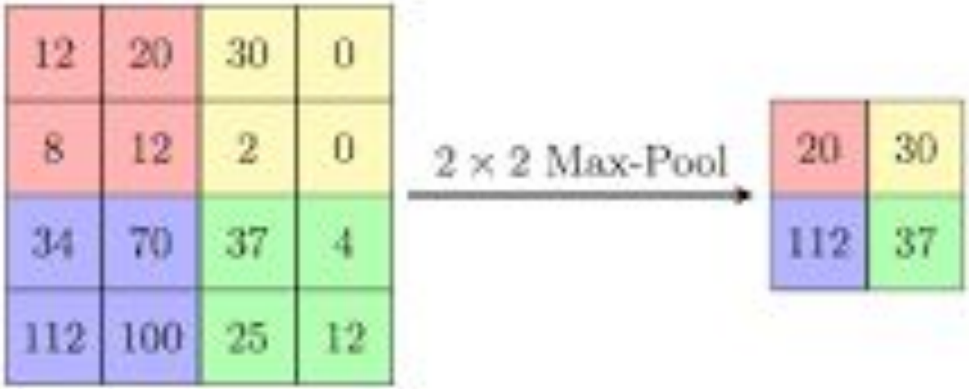
Kernel
(filter)

31			

Output

ARCHITECTURE

- Max Pooling is a pooling operation that calculates the maximum value for patches of a feature map, and uses it to create a downsampled (pooled) feature map.

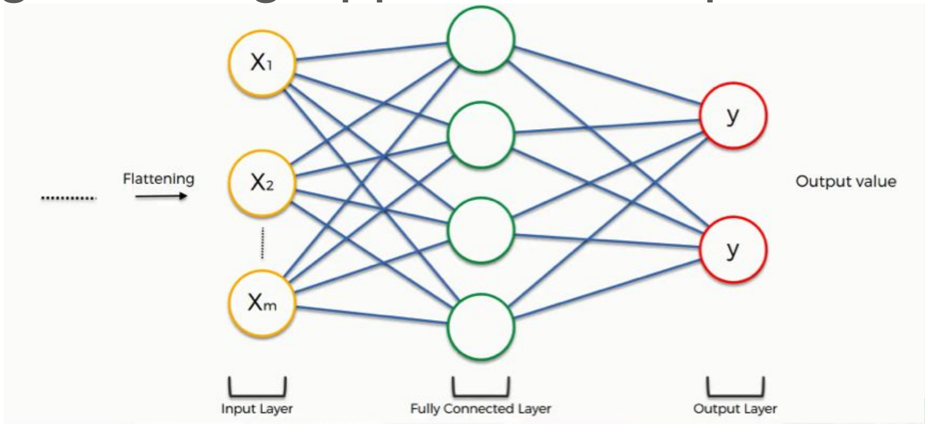


12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

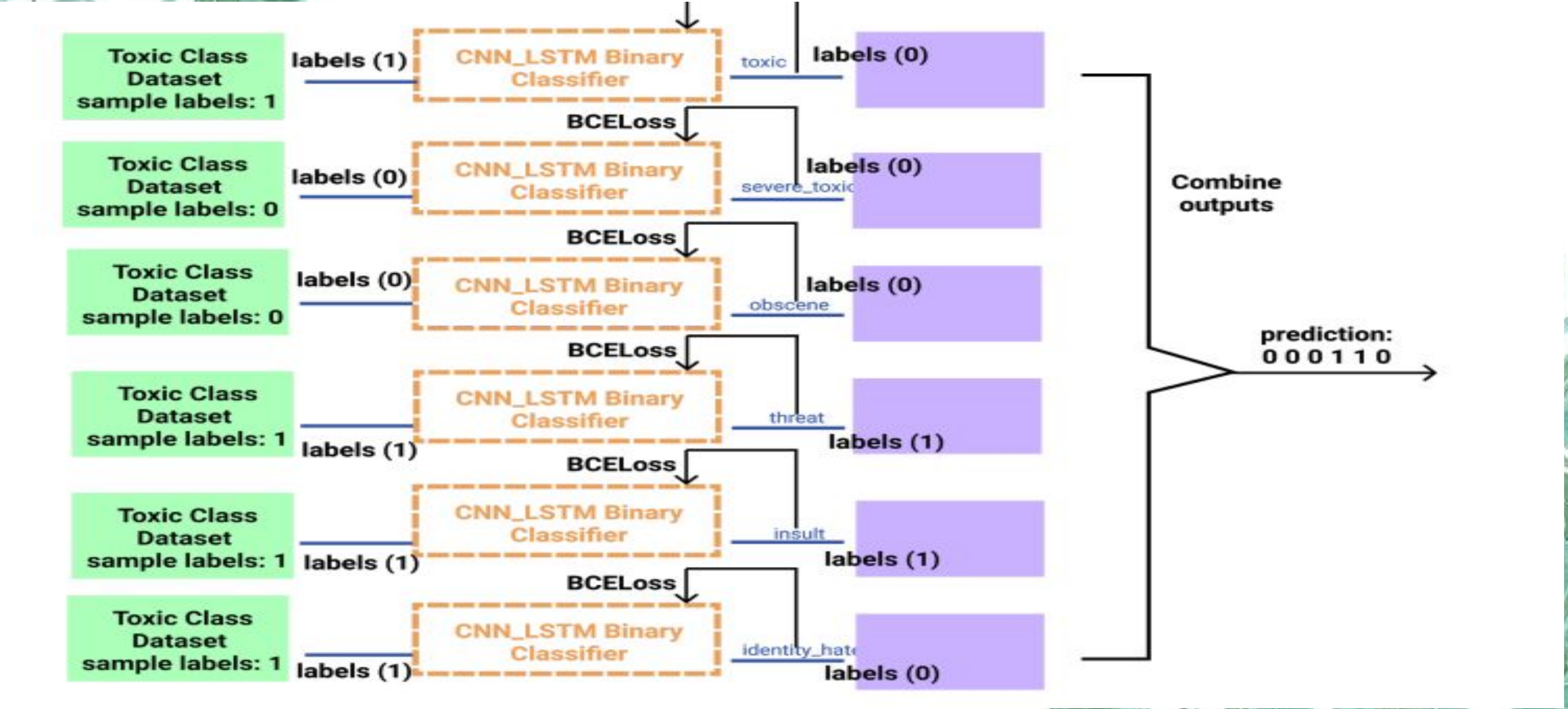
20	30
112	37

ARCHITECTURE

- The Fully Connected (FC) layer consists of the weights and biases along with the neurons and is used to connect the neurons between two different layers.
- The project focuses on multi-labelling classification therefore the following Training Approach is implemented.



ARCHITECTURE


Toxic Class Dataset sample labels: 1
labels (1)
CNN_LSTM Binary Classifier
toxic
labels (0)
BCELoss
Toxic Class Dataset sample labels: 0
labels (0)
CNN_LSTM Binary Classifier
severe_toxic
labels (0)
BCELoss
Toxic Class Dataset sample labels: 0
labels (0)
CNN_LSTM Binary Classifier
obscene
labels (0)
BCELoss
Toxic Class Dataset sample labels: 1
labels (1)
CNN_LSTM Binary Classifier
threat
labels (1)
BCELoss
Toxic Class Dataset sample labels: 1
labels (1)
CNN_LSTM Binary Classifier
insult
labels (1)
BCELoss
Toxic Class Dataset sample labels: 1
labels (1)
CNN_LSTM Binary Classifier
identity_hate
labels (0)
Combine outputs
prediction: 0 0 0 1 1 0

CONCLUSION

The Toxic Comment Classification System using Deep Learning project represents a significant step towards fostering a safer and more respectful online environment. By harnessing the power of deep learning techniques, specifically convolutional neural networks, the project successfully addresses the pressing issue of toxic comments within online discussions.

REFERENCES

[1] Anton Liu , Yue Zhuang, Grace Wu “Toxic Comment Classification”, IEEE journal, May 2022

[2] Hong Fan Wu Du, Abdelghani Dahou, Ahmed A. Ewees, Dalia Yousri, Mohamed Abd Elaziz “Social Media Toxicity Classification Using Deep Learning: Real-World Application” Journal of Emerging Technologies and Innovative Research (JETIR), 1 June 2021.

[3] Victor Blanes Marti, “TOXIC COMMENT CLASSIFICATION USING CONVOLUTIONAL AND RECURRENT NEURAL NETWORKS”,International Research Journal of Modernization in Engineering Technology and Science June 2018.

[4] Danilo Dessì,Diego Reforgiato Recupero and Harald Sack, “An Assessment of Deep Learning Models and Word Embeddings for Toxicity Detection within Online Textual Comments“, International Journal of Engineering Applied Sciences and Technology, 2022 .

[5] Noah Ballinger, “Using a BERT-based Ensemble Network for Abusive Language Detection“, 27th International Conference on International Conference on Machine Learning, 2010.

thank
you