Toxic Comment Detection

1405036
1405045
1405051

1. Introduction

The internet has opened up a Pandora's Box of unforeseen issues. Cyberbullying is one of them which has gained prominence in recent times. Social media is riddled with offensive comments which impact people negatively. Identifying such comments have become of the utmost importance. This project aims to explore various machine learning approaches (both traditional and deep learning) for detecting toxic comments. We study the impact Naïve Bayes (NB), Decision Tree (DT), Logistic Regression (LR), K Nearest Neighbors (K-NN), NB-LR, Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM), Bi-directional Long Short-Term Memory Networks (Bi-LSTM) on identifying toxicity in text. We evaluated our approaches on Wikipedia comments from the Kaggle Toxic Comment Classification dataset.

2. Dataset Description

We collected the dataset from Jigsaw/Google's Toxic Comment Classification Challenge on Kaggle. The dataset contains 159,571 labeled examples of Wikipedia comments that have been labeled by human raters for toxic behavior. The data comes in the schema of <id, commentText, toxic, severeToxic, obscene, threat, insult, identityHate>, where the labels for toxic, severeToxic, obscene, threat, insult, and identityHate are all boolean labels.

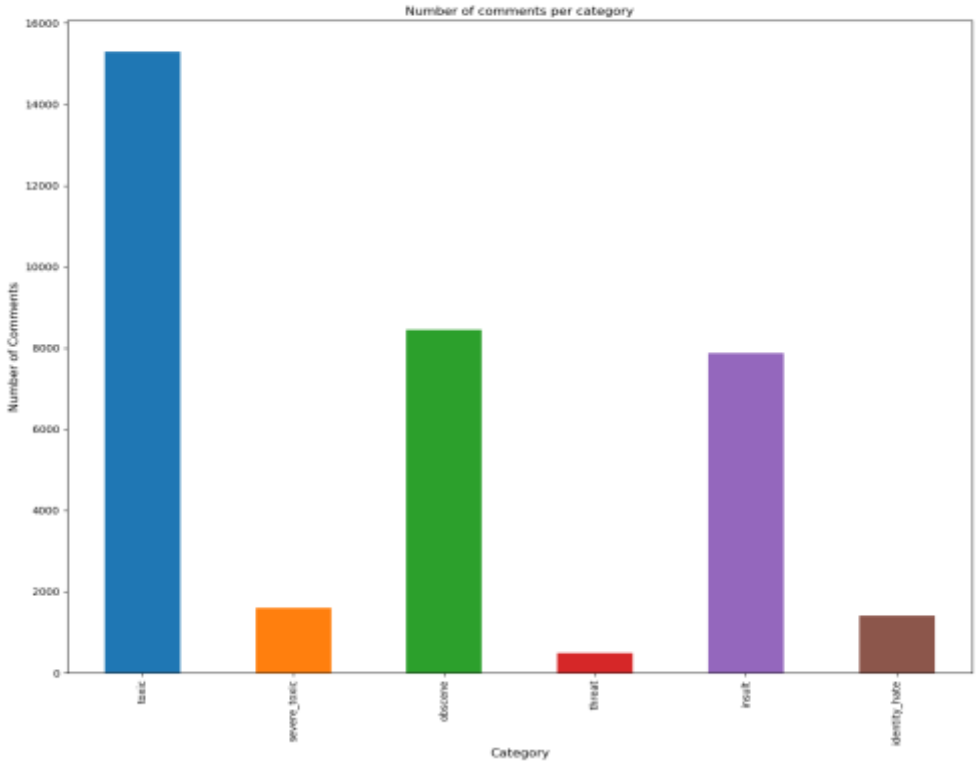

Figure 1 Category Distribution of the dataset

3. Data preprocessing

Before feeding into the models, raw texts of comments required some preprocessing. We removed stop words. We also cleaned our corpus by correcting the spelling of words. We split every text by white-space and removed suffices from words by stemming them with Snowball Stemmer from NLTK library. Finally, we rejoined the word tokens by white-space to present our clean text corpus which had been tokenized later for feeding our models.

Balancing the dataset with subsampling led us to having a total 28981examples for training and 7246 examples for test. For multi-label classification, we directly use the balanced dataset. For binary classification, we consider a comment as toxic if at least one of the labels (toxic, severeToxic, obscene, threat, insult, identityHate) is 1.

4. Performance Metrics

We use accuracy, precision, recall, F1-score to measure the performances of our models. For multi-label classification, we consider label wise performance for each category (toxic, severeToxic, obscene, threat, insult, and identityHate) separately.

5. Results

In this section, we describe performance of our traditional machine learning and neural network based deep learning models.

5.1. Results for Traditional Machine Learning Approaches

Binary Classification:

Model	Feature	Acc	Pre	Rec	F1-Score
LR	Lexical	.78	.78	.78	.78
Decision Tree	Lexical	.71	.71	.71	.71
Naïve Bayes	Bigram	.79	.79	.79	.79
KNN	Empath Features	.62	.64	.62	.62

Multi-label Calssification:

For Toxic:

Model	Feature	Acc	Pre	Rec	F1-Score
NB-LR	n-gram (n=1,2)	.81	.81	.81	.81
SVM (One vs Rest)	Unigram	.88	.88	.88	.88

For Severe Toxic:

Model	Feature	Acc	Pre	Rec	F1-Score
NB-LR	n-gram (n=1,2)	.96	.94	.96	.95
SVM (One vs Rest)	Unigram	.96	.95	.96	.95

For Obscene:

Model	Feature	Acc	Pre	Rec	F1-Score
NB-LR	n-gram (n=1,2)	.88	.88	.88	.88
SVM (One vs Rest)	Unigram	.91	.91	.91	.91

For Threat:

Model	Feature	Acc	Pre	Rec	F1-Score
NB-LR	n-gram (n=1,2)	.99	.98	.99	.98
SVM (One vs Rest)	Unigram	.99	.99	.99	.99

For Insult:

Model	Feature	Acc	Pre	Rec	F1-Score
NB-LR	n-gram (n=1,2)	.86	.85	.86	.85
SVM (One vs Rest)	Unigram	.88	.87	.88	.87

For Identity Hate:

Model	Feature	Acc	Pre	Rec	F1-Score
NB-LR	n-gram (n=1,2)	.96	.95	.96	.95
SVM (One vs Rest)	Unigram	.97	.96	.97	.96

5.2. Results for Deep Learning Apporaches

CNN:

Number of Filter	Filter length	Validation Acc.
128	3	0.8625 [Early Stop at 10 epoch]
256	5	0.8628
512	**5**	**0.8697[Early Stop at 9 epoch]**

Performance on Test:

Accuracy: 0.8666666666666667

Precision: 0.8665851682860687

Recall: 0.8666666666666667

F1 Score: 0.8662863279529587

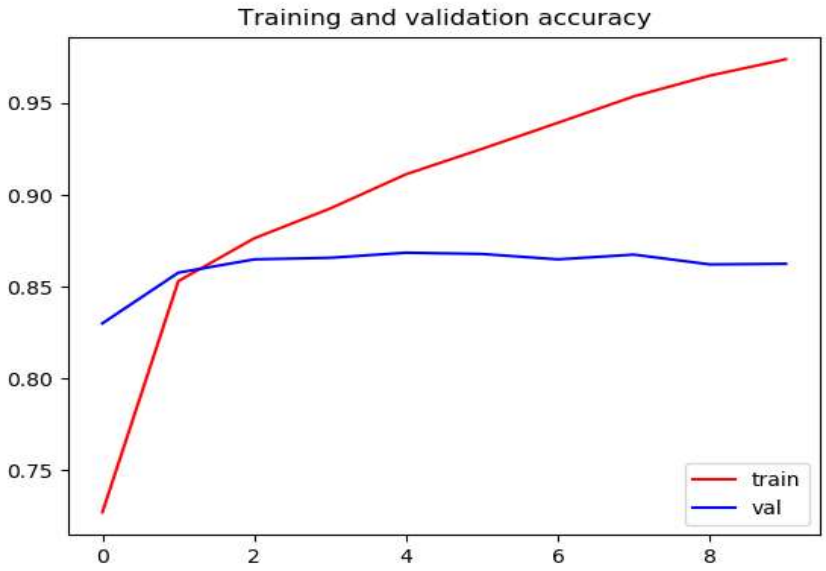

Figure 2 CNN-1

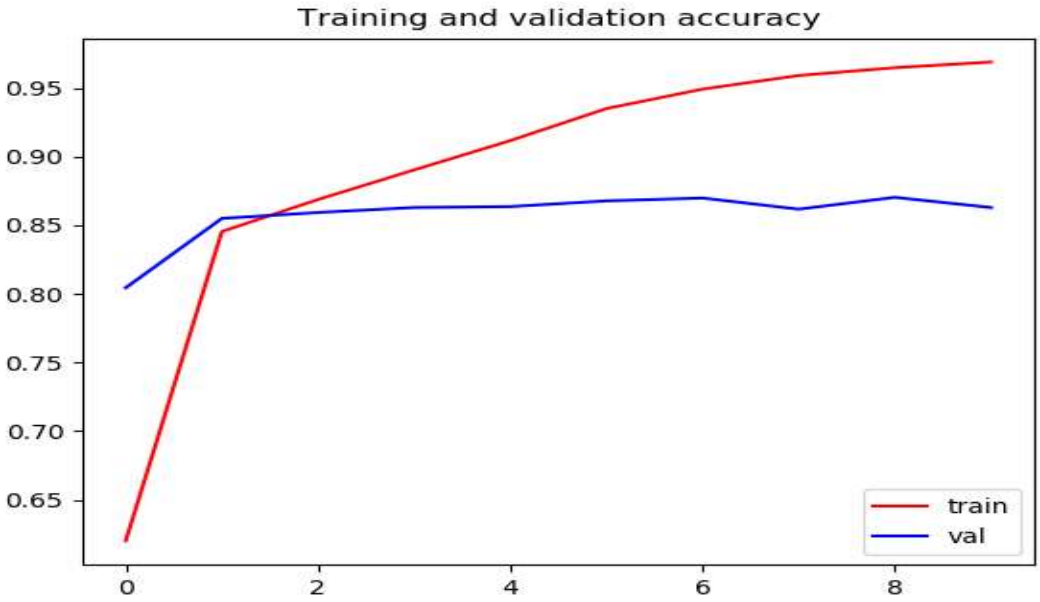

Figure 3 CNN-2

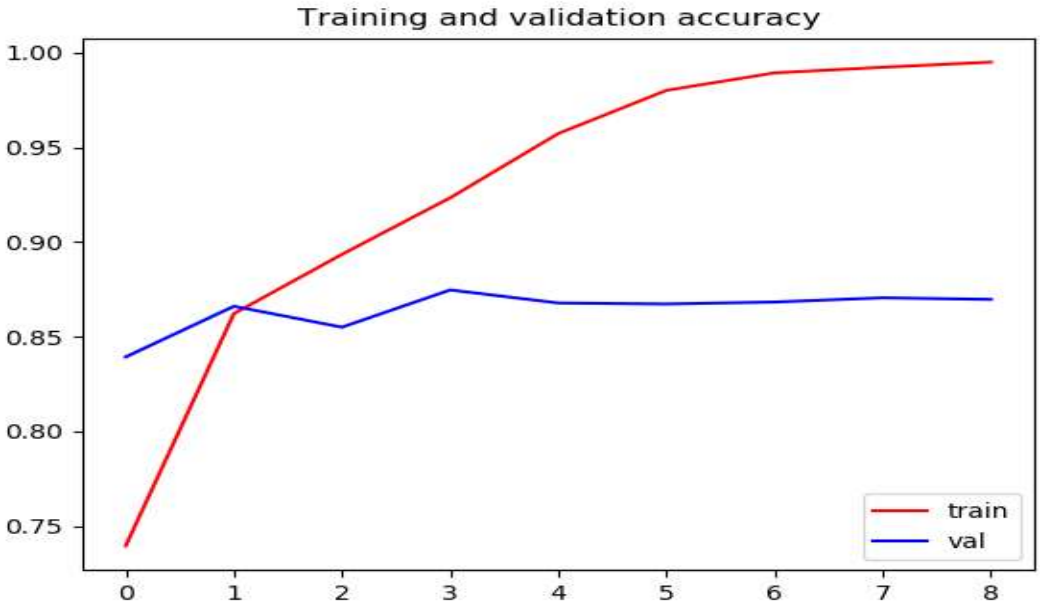

Figure 4 CNN-3

LSTM:

No. of Neuron LSTM	Validation Acc.
50	0.7685 [Early Stop at 08 epoch]

Epoch 00008: ReduceLROnPlateau reducing learning rate to 1.0000000474974514e-05.

Performance on Test:

Accuracy: 0.7696342305037958

Precision: 0.8034147773376381

Recall: 0.7696342305037958

F1 Score: 0.7554179853830427

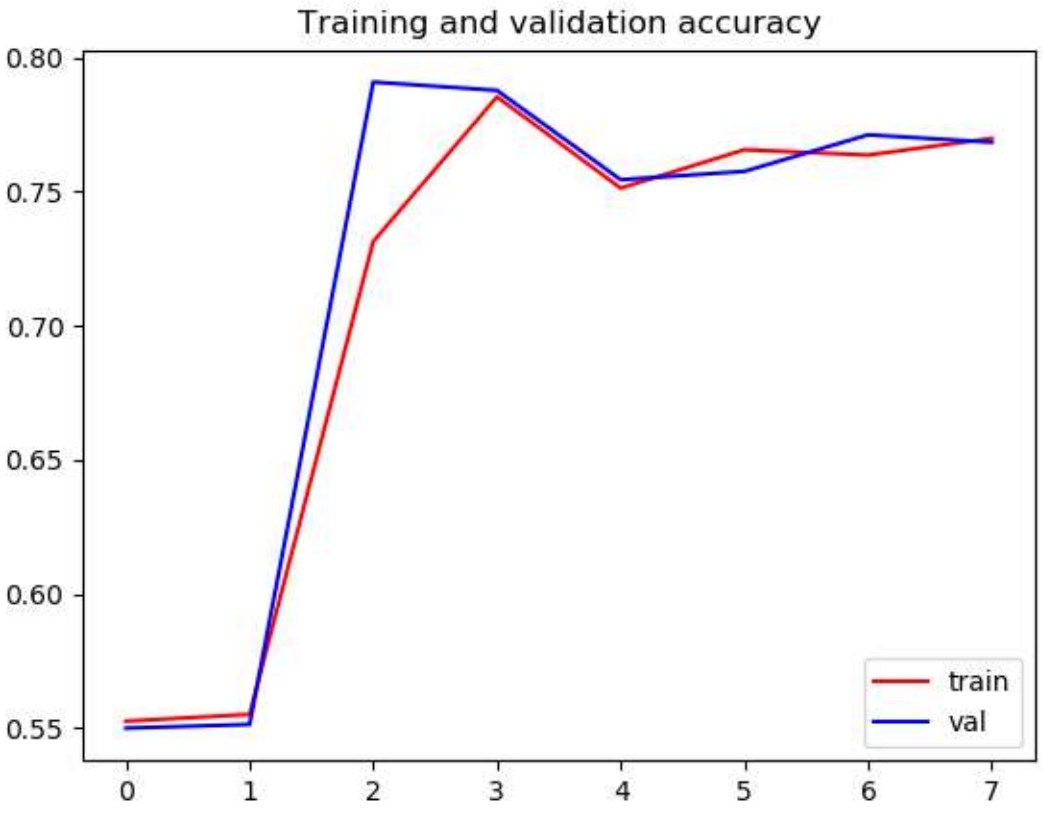

Figure 5 LSTM Accuracy

LSTM IN REVERSE DIRECTION:

No. of Neuron LSTM	Validation Acc.
50	0.8758

Performance on Test:

Accuracy: 0.875224292615597

Precision: 0.8752903295513097

Recall: 0.875224292615597

F1 Score: 0.8748100363858518

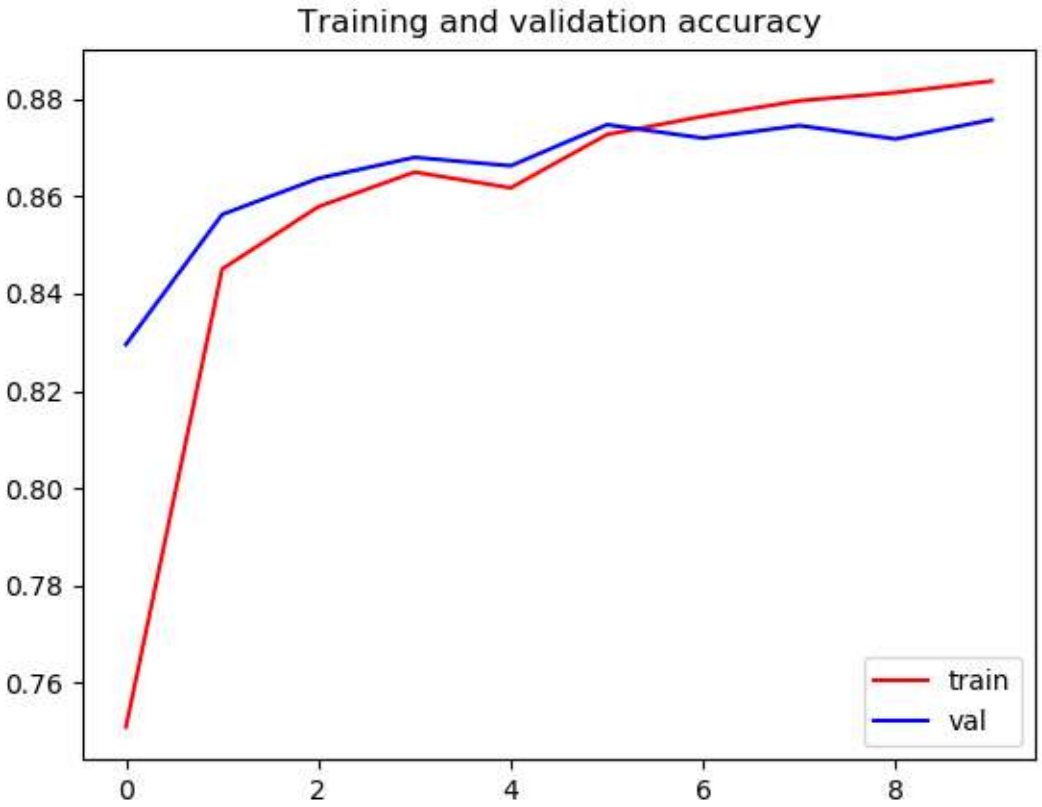

Figure 6 LSTM-Reverse Accuracy

Bi-LSTM:

No. of Neuron Bi-LSTM	Neuron in FC Layer	Validation Acc.
50	64	0.8834
100	**64**	**0.8872**
200	64	0.8849

Performance on Test:

Accuracy: 0.881159420289855

Precision: 0.881799704366389

Recall: 0.881159420289855

F1 Score: 0.8805581973124

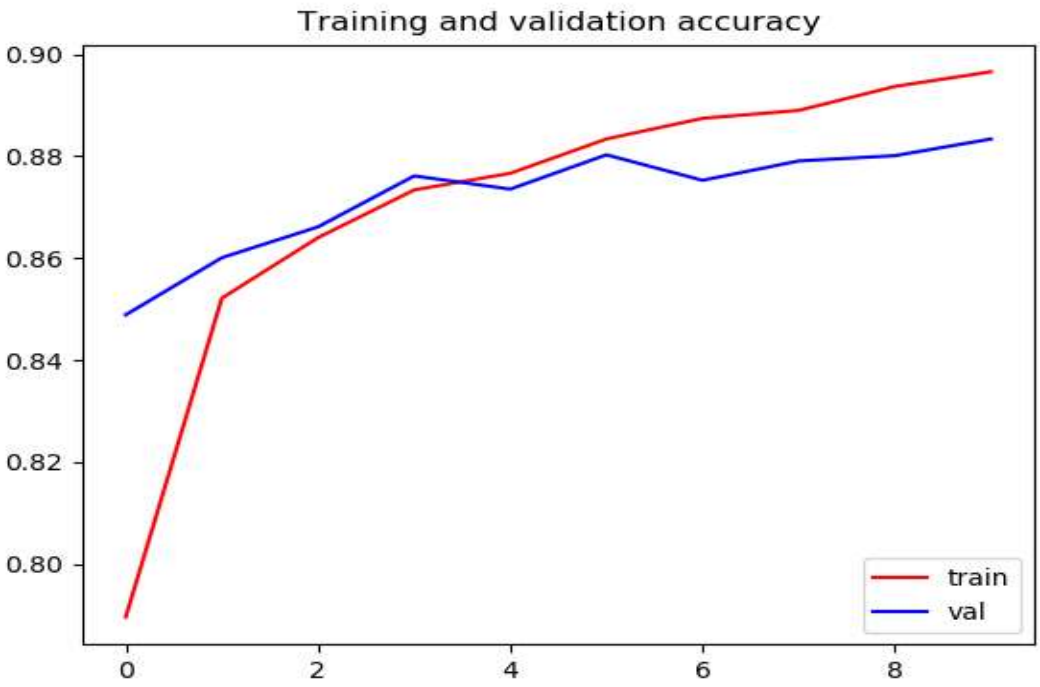

Figure 7 Bi-LSTM 1

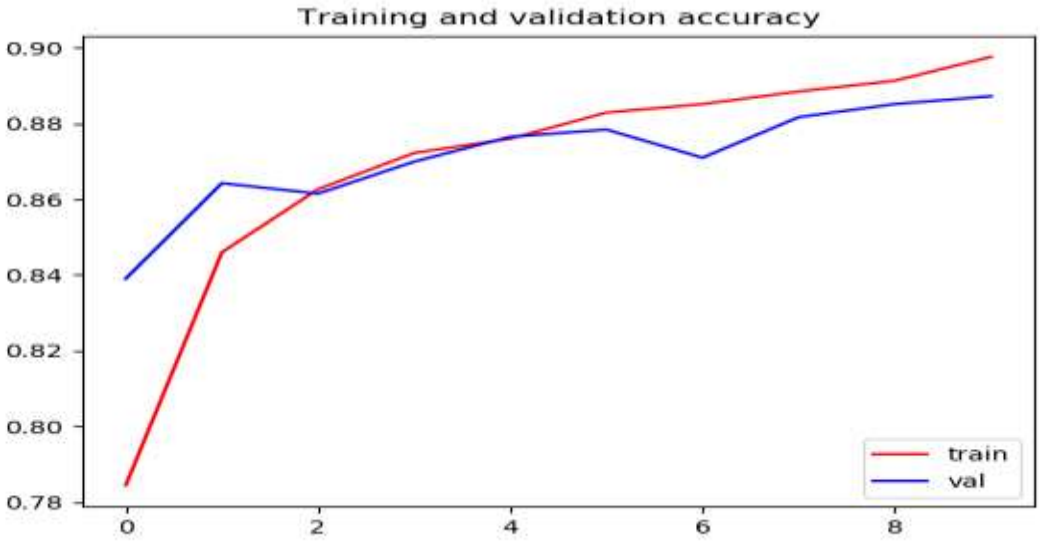

Figure 8 Bi-LSTM 2

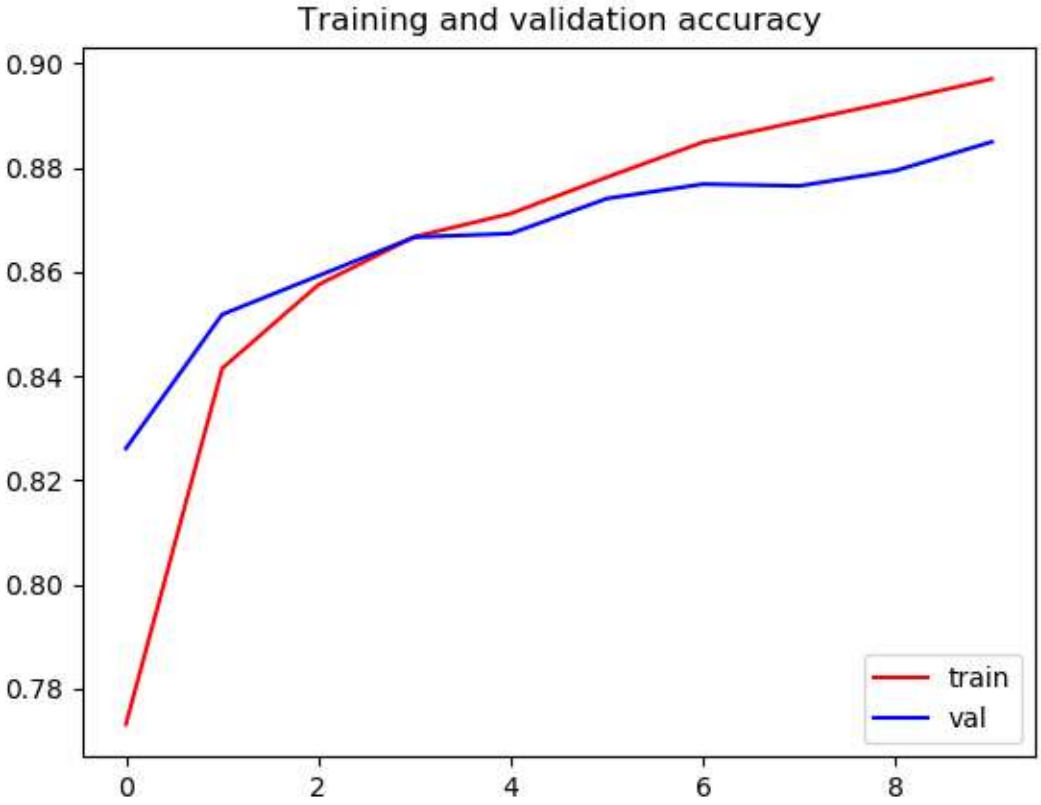

Figure 9 Bi-LSTM 3

6. Best performance Model:

Best Result was observed with Bi-LSTM with 100 Neurons and 64 Neurons in fully connected layer where Accuracy, Precision, Recall and F-1 Score were all above 88%

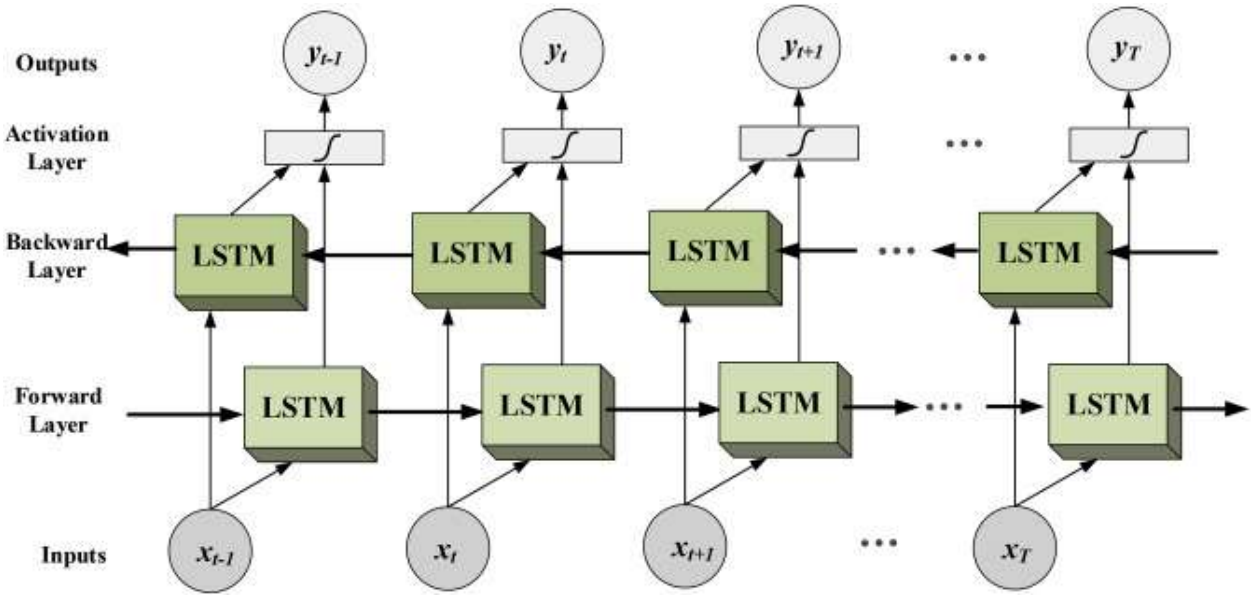

Figure 10 Bi-LSTM Organizationl

Figure 11 Bi-LSTM Model

Conclusion

From our observation, bi-directional LSTM offers the best performance for identifying toxic and offensive comments. This model could be used for screening comments on different social media or possibly as a browser extension to screen comments across all websites.