Introduction to Deep Learning Final Project

Toxic Comment Classification

What problem did you solve?

Goal: detect different types of of toxicity like threats, obscenity, insults, and identity-based hate.

The current models out there are still making errors and they don’t allow users to select which types of toxicity they’re interested in finding.

What ML approach do you use, or what methods does your app use?

Baseline

baseline model consisting of an LSTM layer followed by a global max pooling layer along with two dense layers. achieved a validation accuracy of around 0.79 but had low precision and recall for detecting toxic comments.

Class Weights

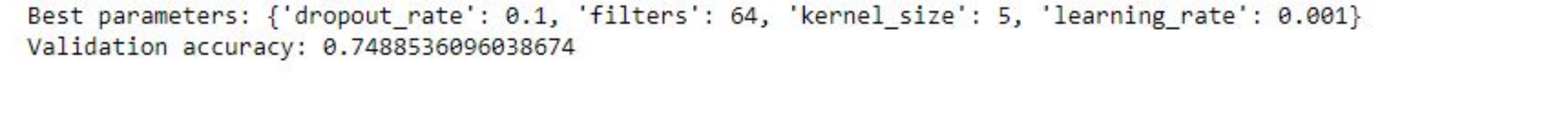
helped improve the precision and recall for detecting toxic comments. However, it did not improve the overall performance of the model by much.

Grid Search

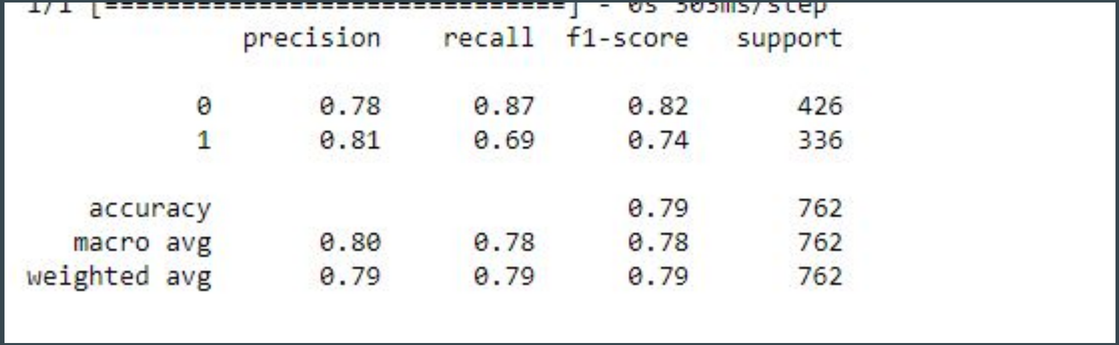
Best parameters: {'dropout_rate': 0.1, 'filters': 64, 'kernel_size': 5, 'learning_rate': 0.001}

Results

Findings



```
Best parameters: {'dropout_rate': 0.1, 'filters': 64, 'kernel_size': 5, 'learning_rate': 0.001}
Validation accuracy: 0.7488536096038674
```



```
1/1 [==============================] - 0s 305ms/step
              precision    recall  f1-score   support

           0       0.78      0.87      0.82       426
           1       0.81      0.69      0.74       336

    accuracy                           0.79       762
   macro avg       0.80      0.78      0.78       762
weighted avg       0.79      0.79      0.79       762
```

Conclusion:

Overall In this project I built a multi-headed model using Keras that can detect different types of toxicity in text data. I used a dataset of Wikipedia comments that had been labeled for toxicity by human raters.

I first began by performing exploratory data analysis and cleaning the data to prepare it for modeling. I then trained a baseline model consisting of an LSTM layer followed by a global max pooling layer along with two dense layers. The model achieved a validation accuracy of around 0.79 but had low precision and recall for detecting toxic comments.