# Results

## Baseline Results

The base model for performance comparison is the SVM model and it showed promising accuracy of 94.13%. The base model showed promising results, easy to implement and had fast runtime. The following models will be compared to the SVM model’s performance. For better readability I will refer to the racism class as the positive class of tweets containing racist language and the non-racism class as the negative class of non-hateful tweets. The resulting classification report of the SVM model shows that the SVM model precision for the positive class is 0.96 compared to 0.93 for the negative class but the recall is much higher for the negative class with 0.97 compared to a recall of 0.9 for the positive class. This shows that the SVM model is better at predicting racist tweets but is able to better detect non-racist tweets.

Classification Scores and corresponding Confusion matrix

A picture containing table

Description automatically generated

Chart, treemap chart

Description automatically generated

The first model implemented after the SVM model was the Bertbase MLP model which had a higher accuracy score of 95.18 %. However when plotting the loss and accuracy graphs of the MLP model the graphs below show a high degree of over fitting despite hyperparameter tuning efforts. The original implementation suggested using a random initializer but replacing the initializer with the HeNormal initializer reduced the overfitting. Moreover, decreasing the neurons in the dense layer further reduced the overfitting of the model. This did not seem to impact the accuracy of the test dataset but such a large gap between training and validation loss would hinder its generalization and suggests implementing further models.

Chart, line chart

Description automatically generated

BertBaseMLP training plots

### Bertbase CNN Model Experimentation and Results

The Bertbase CNN model required much more effort in hyperparameter tuning and experimenting with different architectures and layers to achieve final satisfactory results. As previously discussed, the following layers were used to experiment which produces best performance for detecting racism tweets: pooling, sequence, and concatenated encoder layers. The following configurations were essential for the originally BertCNN model improvement regardless of the layer tested: drop out layers, number of neurons chosen for dropout layer, number of neurons in the Dense Layer with the results shown below. The optimal parameters were chosen based on accuracy and validation loss results. The model performs very well on the training set with a consistently high training accuracy and an ideal training loss close to zero. The initially inspired architecture with 512 neurons in the Dense layer and one dropout layer with a rate of 0.1 showed promising accuracy but seemed to largely overfit with a validation loss of 0.42. Therefore, another dropout layer with a rate of 0.2 was added which slightly decreased the loss. However, decreasing the number of neurons in the Dense layer significantly decreased the validation loss and narrowed the gap between the training and validation loss. An observation made from the results in the table below was the trade-off between dropout layers and the number of neurons in the dense layer. The model’s accuracy generally increased after adding dropout layers and increasing the rate. Moreover, adding a dropout layer between the outputs from the BERT model and the first convolutional layer in the dense layer decreased the validation loss and increased the validation accuracy thus implying an important role in the model’s architecture that was not previously suggested. Moreover, decreasing the number of neurons in the dense layer significantly decreased the validation loss suggesting a less complex model is more suitable for this task. However, after increasing the dropout layer and the rate the validation loss seems to slightly increase but also offers a higher accuracy. Therefore, the chosen hyperparameters were a balance of the tradeoff and chose a low validation loss but preferred higher accuracy model. The final model architecture that was applied to all the following variations of the BERTCNN model was: 3 dropout layers with a rate of 0.4 with a higher rate of 0.6 for the first dropout layer and 128 neurons for the Dense layer which gave an accuracy of 96.3% and validation loss of 0.17.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dropout Layers # | Dropout neurons | Dense Layer neurons | Test Accuracy | Validation Loss |
| 1 | 0.1 | 512 | 95.43 | 0.42 |
| 2 | 0.2 | 512  256 | 95.46  95.15 | 0.36  0.194 |
| 3 | 0.2  0.4 | 256 | 95.2  95.8 | 0.22  0.185 |
| 3 | 0.2  0.4  0.6 | 128 | 95.8  96.3  96.4 | 0.21  0.17  0.24 |
| 4 | 0.4 | 128 | 94.9 | 0.26 |

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedTable 5.1: Test Accuracies of Models

Example of Results Before and After Tuning

Comparison of the three key layers' performance:

The three layers showed very close accuracies where the sequence layer showed the highest accuracy of 96.5, followed by the pooling layer with 96.2, and finally the encoder layer with 96.0. Therefore, classification scores were produced to further analyze the variations of the model. Notably, the encoder layer scored much lower than the sequence and pooled outputs on recall. The encoder layer has a high precision of 0.94 at classifying the positive class but struggles to detect them with a recall of 0.88 compared to the pooled and sequence layers. The pooled layer only slightly shows better performance than the sequence layer with 1% higher scores of precision, recall, and F1 score. Another factor that contributes to the decision of using the pooled output was the lower validation loss compared to the sequence layer that was observed while running the models.

Encoder layers

precision recall f1-score

0 0.95 0.95 0.95

1 0.94 0.88 0.91

Sequence layer

precision recall f1-score

0 0.97 0.95 0.96

1 0.93 0.93 0.93

Pooled output

precision recall f1-score

0 0.97 0.95 0.97

1 0.95 0.94 0.94

### GCRNN Model Experimentation and Results

The proposed GCRNN model inspired by Lee et al. reported 97% accuracy results on their racism dataset and 95% on a dataset similar to the one used in this study. When initially applied to this dataset, an accuracy of 90.56% was achieved which was much lower than reported results and even the baseline model. Therefore, a few adjustments were experimented to test if the accuracy would increase. The following hyperparameters were tested: input\_length, dense layer neurons, number of epochs, and batch size.

After the final hyperparameters were decided a further analysis was done by producing the classification scores for the GRCNN model. The GRCNN model has the lowest precision score thus far but notably the highest recall score. It seems that the model can detect the positive class very well with a recall of 0.97.

Precision recall f1-score

0 0.96 0.93 0.95

1. 0.91 0.97 0.93

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Input Length | epochs | Dense Layer neurons | Batch Size | Test Accuracy |
| 5 | 100 | 16 | 16 | 90.56 |
| 10 | 60  40 | 16 | 16  32 | 0.78  0.80 |
| 15 | 40 | 16  32  128 | 32 | 93.83  94.07  94.18 |
| 20 | 40 | 128 | 32 | 0.76 |

### 5.4 Models Performance and Generalizability of Ukraine and Afghanistan Datasets

The Ukraine and Afghanistan twitter datasets contain text that is relevant to the time of the crisis and contains public opinions on immigration and racism as the events were ongoing. The tweets were labelled using the TextBlob library with negative, positive, and negative. The distribution of the labels is shown in the pie charts below. It is noticeable that the Afghanistan tweets show a higher percentage of negative tweets of 53% where nearly half of the tweets show negative sentiment concerning immigration. Meanwhile, the Ukraine tweets show a 39.1% negative sentiment when immigration is discussed. The GRCNN and the best Bertbase CNN model were used to make predictions on both datasets. This study is interested if the models can confirm the hypothesis that there is a larger anti-immigration sentiment in the Afghanistan crisis tweets compared to the Ukraine crisis by detecting a notably higher number of racist tweets. Moreover, the study is interested in how well these models perform on these external corpuses thus commenting on the generalizability of the models.

The BertBase CNN on the earlier dataset demonstrated an accuracy of 96.3% and recall of 0.94. Meanwhile the GCRNN model demonstrated an accuracy of 94.2% and a recall of 0.97. Recall is important to this analysis because it will provide a measure of how good the model is at detecting racism tweets. Predictions were run on both datasets and a few measures were considered: percentage of racist tweets predicted, Recall, and F1 score. The GRCNN model detected a very similar percentage of racist tweets compared to the actual percentage of racist tweets. However, the low recall and F1 score a high misclassification rate despite a representative prediction. The GRCNN however shows a much higher recall for racist tweets in the Afghanistan dataset than the Ukraine dataset which will be discussed further later. The BERTCNN performed much better on the Ukraine dataset than on the Afghanistan Dataset. The BERTCNN only predicted half of the actual percentage of racist tweets and predicted 8% more racist tweets in the Ukraine dataset. The GRCNN model clearly performed much better on the Afghanistan tweets while the BertBase Model performed slightly better on the Ukraine tweets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Model | Actual Racist tweets | Predicted Racist tweets | Recall | F1 score |
| Ukraine | GCR-NN | 39.1% | 30.8% | 0.51 | 0.43 |
| Afghanistan | GCR-NN | 53% | 40 % | 0.70 | 0.56 |
| Ukraine | BERTCNN | 39.1% | 47% | 0.57 | 0.53 |
| Afghanistan | BERTCNN | 53% | 25.9% | 0.3 | 0.40 |

Chart, treemap chart

Description automatically generatedChart, treemap chart

Description automatically generated

Chart, treemap chart

Description automatically generatedChart, treemap chart

Description automatically generatedGCRNN Afghan Vs Ukraine

BertBase Afghan Vs Ukraine

Further Analysis:

Further Analysis:

The BertCNN model showed a noteworthy poor performance on the Afghan Dataset. Therefore, the model predictions were compared to the actual values and the VADER compound score as shown in the tables below.

Text

Description automatically generated with medium confidence

Afghanistan Dataset

Limitations of the VADER library in this study:

The VADER library is tasked to allocate a negative score for negative sentiment found in the dataset. Therefore, many tweets are misclassified as racist due to the negative sentiment present in the tweet such as the last tweet where there are many negative words but the tweet itself was not racist which is highly misleading to the algorithm. Therefore, tweets were relabeled as negative only if they are associated with a high negative score. After observing more tweets, the majority of the tweets scoring above -0.5 were correctly labelled as racist. Consequently, this reduced the percentage of racist tweets in the corpus to 36%.

Results after adjusting the VADER scoring:

The adjustment to the dataset Afghanistan BERT model improved recall to 0.36 and F1 score to 0.41 and predicted a 25% of racist tweets. This estimate is much closer to the actual percentage than the previous prediction of the previous implementation. The same strict labelling was applied to the Ukraine dataset but with a small difference in implementation. This also decreased the percentage of false positives from 39.62% to 23.50%. The Ukraine dataset negative tweets dropped to 21% which would already impact the model’s performance due to imbalance. Therefore, for better analysis, the neutral tweets were dropped which balanced the dataset between negative and positive tweets in the Ukraine dataset. There was a slight increase in recall from 0.51 to 0.58 and the F1 score from 0.43 to 0.54.

As the GCRNN model showed the most promising results in both datasets a further investigation of the model’s sensitivity is also measured by the precision value to better understand the difference in performance in both datasets. The precision for the Ukraine dataset was much higher than the Afghanistan dataset with 0.5 compared to 0.38. Therefore, a word cloud of the correctly predicted racist tweets was generated for both datasets. As shown below, the Afghanistan dataset word cloud for the correctly classified racist tweets contains a lot of hateful words. Most Notably the words: ”immigrant, Taliban, illegal, terrorist, and anti, refugees”, seem to be associated with the Afghanistan immigration tweets. On the other hand, the word cloud produced for the Ukraine tweets has a high mention of some notable words: “ racism, racist, anti, black, and african”. As the original dataset was trained on generally racist tweets this could explain why the model had higher precision and sensitivity to the words in the Ukraine dataset rather than the Afghanistan dataset. This could be correlated to the nature of the words detected and present in the Ukraine corpus which is more relevant to racism while the Afghanistan tweets while also racist use a more specific anti-immigration language which is not present in the general corpus.

Text

Description automatically generated

Text

Description automatically generated

|  |  |
| --- | --- |
| Model | Test Accuracy |
| SVM | 94.13 |
| Bertbase MLP | 95.18 |
| BERTbase CNN:Pooling Layer | 95.8 |
| BERTbase CNN:Sequence Layer | 96.5 |
| BERTbase CNN:Encoder Layer | 96.0 |
| GRU-NN | 94.14 |