COL774 Assignment 2 Report

Dhananjay Sapawat

7 October 2023

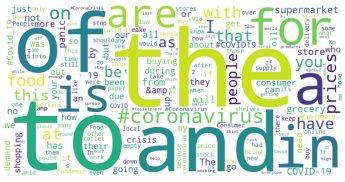
1 Text Classification

1.1 Naive Bayes Multiclass

1.1.1 Accuracy of Algorithm

Training Set Accuracy: 85.16%
Validation Set Accuracy: 66.38%

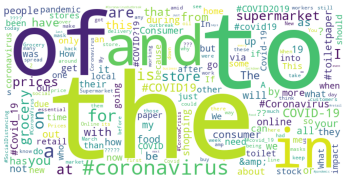
1.1.2 Word Clouds



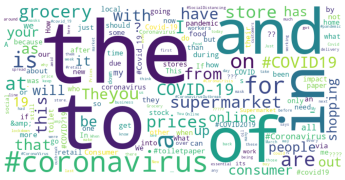
(a) Negative Class (Training)

(b) Negative Class (Validation)

Figure 1: Word Clouds for the Negative Class



(a) Neutral Class (Training)



(b) Neutral Class (Validation)

Figure 2: Word Clouds for the Neutral Class

(a) Positive Class (Training)

(b) Positive Class (Validation)

Figure 3: Word Clouds for the Positive Class

1.2 Random and Positive Prediction

1.2.1 Training Set

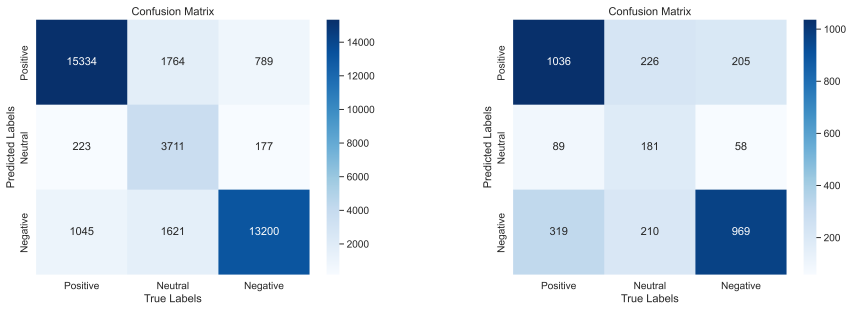
- Random Accuracy: 33.29%
- Improvement over Random: 51.87%
- Positive Accuracy: 43.85%
- Improvement over Positive: 41.31%

1.2.2 Validation Set

- Random Accuracy: 32.52%
- Improvement over Random: 33.86%
- Positive Accuracy: 43.85%
- Improvement over Positive: 22.53%

1.3 Confusion Matrix

1.3.1 Algorithm Model

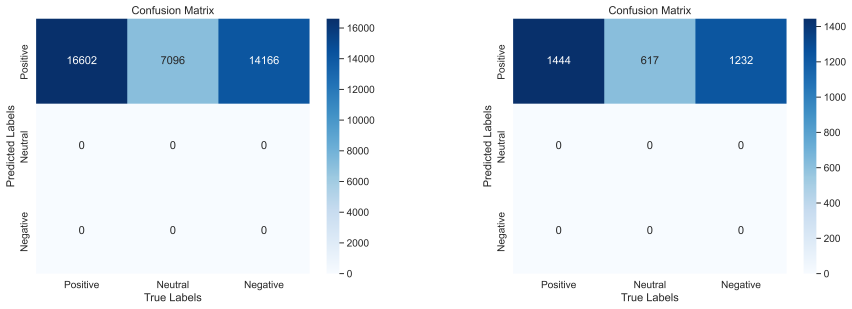


(a) Training Data (b) Validation Data

Figure 4: Confusion Matrix for Algorithm Model

In both training and validation data confusion matrices for the Algorithm Model, the highest diagonal entry is in the category "Positive". This means that the model is performing well in correctly identifying instances of the "Positive" category, indicating a high true positive rate for "Positive".

1.3.2 Positive Model

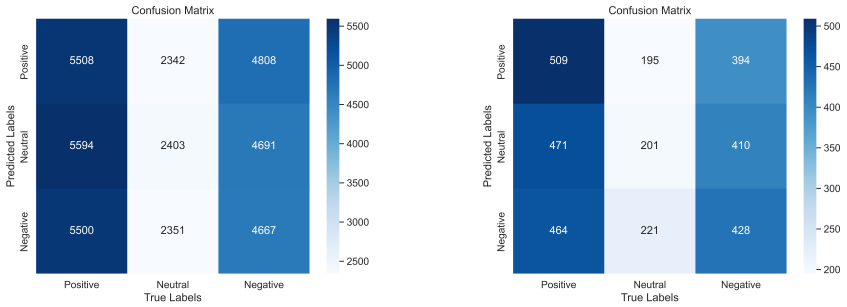


(a) Training Data (b) Validation Data

Figure 5: Confusion Matrix for Positive Model

In both training and validation data confusion matrices for the Algorithm Model, the highest diagonal entry is in the category "Positive". This means that the model is performing well in correctly identifying instances of the "Positive" category, indicating a high true positive rate for "Positive".

1.3.3 Random Model



(a) Training Data (b) Validation Data

Figure 6: Confusion Matrix for Random Model

In both training and validation data confusion matrices for the Algorithm Model, the highest diagonal entry is in the category "Positive". This means that the model is performing well in correctly identifying instances of the "Positive" category, indicating a high true positive rate for "Positive".

1.4 Lemmatizing and Stop-words

1.4.1 Word Clouds

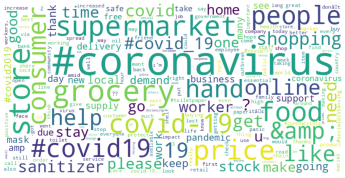
(a) Negative Class (Training) (b) Negative Class (Validation)

Figure 7: Word Clouds for the Negative Class

(a) Neutral Class (Training)

(b) Neutral Class (Validation)

Figure 8: Word Clouds for the Neutral Class



(a) Positive Class (Training)

(b) Positive Class (Validation)

Figure 9: Word Clouds for the Positive Class

1.4.2 Accuracy of Algorithm with Lemmatizing and Removing Stopwords

Training Set Accuracy: 87.37%
Validation Set Accuracy: 68.54%

1.4.3 Accuracy Change

For the Training Set:

$$\text{Accuracy Change} = 87.37\% - 85.16\% = 2.21\%$$

For the Validation Set:

$$\text{Accuracy Change} = 68.54\% - 66.38\% = 2.16\%$$

1.6 Domain adaptation

1.6.1 Accuracy Results for Domain Adaptation

Subset Size	Domain Adaptation	Without Domain Adaptation	Accuracy Improvement
1%	50.03%	41.48%	8.55%
2%	50.76%	41.02%	9.74%
5%	51.76%	43.77%	8.99%
10%	52.58%	48.38%	4.20%
25%	51.69%	47.35%	4.34%
50%	51.92%	48.54%	3.38%
100%	56.96%	54.27%	2.70%

Table 1: Accuracy Comparison with and without Domain Adaptation

1.6.2 Graph for validation set accuracy

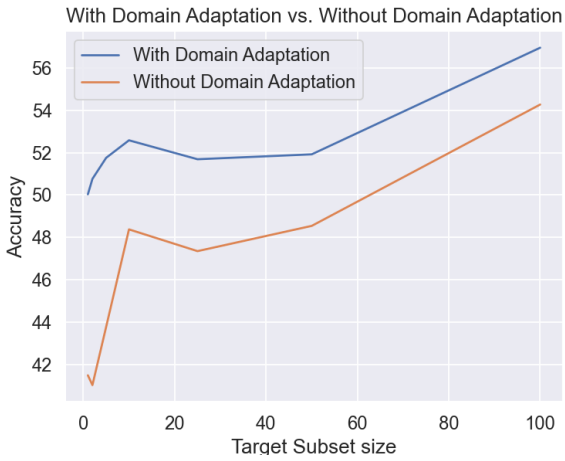


Figure 10: Validation set accuracy for domain adaptation with and without source.

1.6.3 Observations

In the plot shown in Figure 10, we observe that as we increase the size of the target training data, the accuracy of both algorithms generally improves. However, the model trained with source data consistently outperforms the model trained without source data across all Subset sizes. This suggests that domain adaptation, by incorporating source domain knowledge, can lead to better performance on the target domain.