University of Sheffield

COM3110: Sentiment Analysis Assignment Report

Callum Clear

Department of Computer Science

December 9, 2021

1 Introduction

This report discusses a Naïve Bayes sentiment analysis implementation for movie reviews across several configurations and class formats. Different feature selections (words, adjectives, and stop words), as well as intensifier and negation words, are considered. It will begin by describing each implementation and follow by testing and discussion of results.

2 Implementation

Preprocessing is completed on the data before feeding to the model, composed of tokenisation and lower casing. The training data is computed to gain prior probability (the probability of each sentiment class) and likelihood (the probability each feature is of each sentiment class) values. Hereby the Naïve Bayes classifier can be calculated for the development and testing data. Once the same preprocessing and feature selection methods are applied, the probabilities of each review's sentiment class are determined. Per class, this is established by multiplying the prior probability and each features likelihood value. The review is then classified by the class possessing the highest probability.

The other configurations function as follows:

Adjectives - An adjective word-list is read into the model and used as a filter to solely select adjectives as features.

Stop words - A stop word list is read into the model and used as a filter to remove stop words from features.

Intensifiers - An intensifier word list (words such as 'very') is read into the model. During classification, the previous feature is stored, if it contains an intensifier word then the current features likelihood values are scaled by a factor based on the ordering of its probabilities.

Negation - A negation word list (words such as 'not') is read into the model. During the classification, the previous feature is stored, if it contains a negation word then the current features likelihood values are flipped.

3-Class - To create a 3-Class configuration the 5-Class labels were mapped correspondingly: $f(0|1) = 0$, $f(2) = 1$, and $f(3|4) = 2$.

3 Development

As the development data contains labels, it allows the models to be analysed by obtaining a confusion matrix, accuracy score, and a macro-F1 score. Firstly, the confusion matrix is created by iteration over each review in the development data and incrementing the corresponding grid cell in the matrix (whereby the x and y-axis represent predicted and true labels respectively). The accuracy score can then be computed as the sum of the true positive values divided by the total count to give a percentage. Finally, the macro-F1 score is defined by the F1-measure of each class. This is calculated by dividing the product of two and the true positive value by the sum of itself and the false positive and false negative values (which can be extracted from the confusion matrix).

Class Size	Measure	Majority	Base	Adjectives	Stop words	Intensity	Negation
5	Accuracy	0.283	0.422	0.304	0.400	0.384	0.424
	F-1	0.088	0.314	0.176	0.307	0.258	0.317
3	Accuracy	0.433	0.655	0.470	0.655	0.600	0.659
	F-1	0.201	0.497	0.302	0.508	0.450	0.500

Table 1: Accuracy and F1-measure for class sizes 3 and 5 other different configurations

Table 1 illustrates results obtained over each configuration for a five and three-class model. A majority classifier (classification only considering prior probability) is also tabulated here to demonstrate a baseline performance measure. The macro-F1 scores have also been represented on a graph (see fig 1).

4 Conclusion

When comparing each configuration to the majority classifier it is clear the Naïve Bayes approach was successful at classification to some degree with both a five and three-class structure. While most of the configurations appear

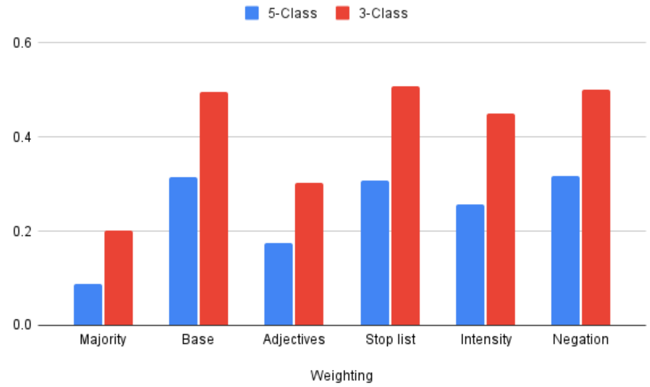

Figure 1: A graph to show the macro-F1 scores of each configuration among two different class sizes

to result in similar scores, it is clear using adjectives as features negatively affects the accuracy as it achieved worse than the base configuration. This is likely due to adjectives being limited features, as they remove verbs and nouns which may correlate to sentiment. The other feature selection method used, stop words, performed much better than using adjectives. The comparison to the base configuration shows that performance was decreased when considering five classes (macro-F1 scores reduced to 0.400 from 0.422) inferring the list used is not be appropriate for sentiment analysis as it is removing useful features. However, a slight increase was seen when only three classes were used (0.508 compared to 0.497) making this configuration the highest-scoring among the 3-Class models. It is unclear why this disparity occurs but when analysing the confusion matrices it seems the stop list results in more neutral classification which is less beneficial when using a larger class size. Overall the span of results when using different features suggest the significance of feature selection when applying Naïve Bayes for sentiment analysis.

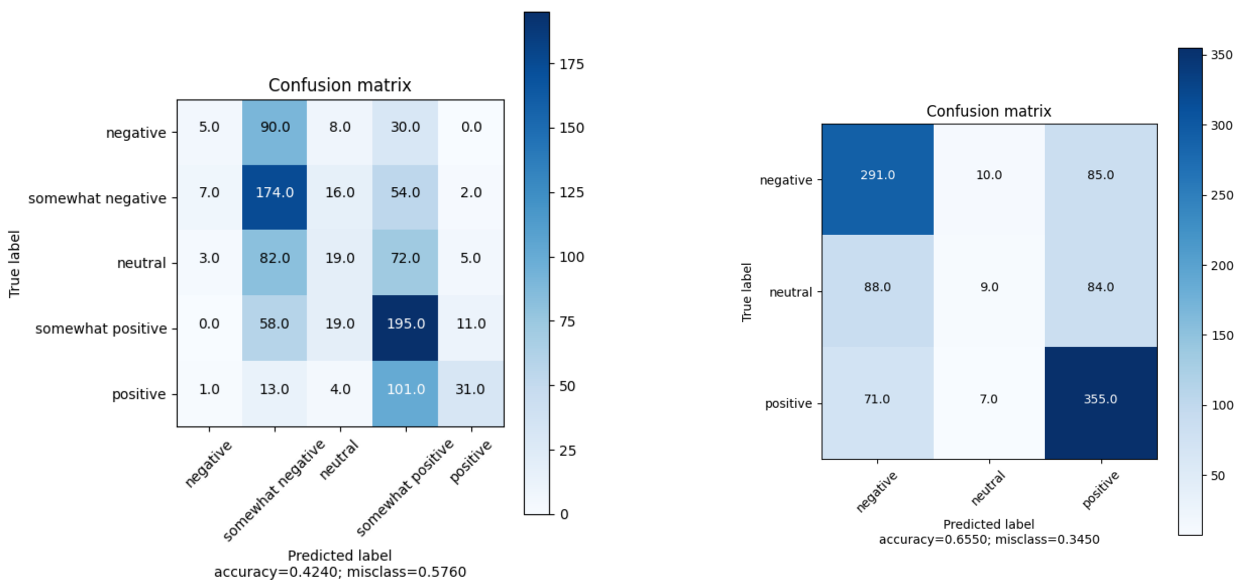

Figure 2: Confusion matrix for three and five class results with the top performing models (Negation and Stop words respectively)

The most accurate feature selection, all words, emerged with respectable scores. Configurations using word intensity and negation worked off these features, showed a small variation of results. Considering word intensifiers was unsuccessful in increasing the scores, which can be explained by studying the confusion matrix for this model; it indicates that far more predictions for the 'somewhat positive' class are being made due to the scaling factor. This may be due to the intensifying scale being too large and thus over-increases the likelihood values for strengthened words. Negation was slightly stronger across both class sizes, scoring the highest of the 5-Class results. However, the development seen was still minute. An improvement to these two methods would be to consider the intensity and negation word list during the training stage.