**Homework 8**

By Shivani Sanjay Mahaddalkar

**Lie Detection and Sentiment Analysis**

**Data Preprocessing**

There are three attributes, namely, lie, sentiment and review. The filter string to word vector is applied to vectorize the string present in review column. It is found under unsupervised filters. Using NGram Tokenizer which will individually tokenize the words. Setting the IDFTransform and TFTransform parameters to False and Lowercase token is set to true so that all upper and lower case words are not treated differently. Setting min term frequency to 1.

Applying the above curated filter on the review column tokenizes the column.

**SVM**

Running linear SVM for lie detection and sentiment analysis, with a training split of 66%. The following are the results.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Overall Accuracy** | **Precision in category I** | **Recall in category I** | **Precision in category II** | **Recall in category II** |
| **Lie Detection** | 45.16% | 0.4 | 0.615 | 0.545 | 0.333 |
| **Sentiment** | 74.19% | 0.769 | 0.813 | 0.722 | 0.813 |

Weka Output:

Table

Description automatically generated

Table

Description automatically generated

**Multinomial Naïve Bayes**

On running the Multinomial Naïve Bayes Text model, with a training split of 66% for the filter defined in the data processing window, the output is as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Overall Accuracy** | **Precision in category I** | **Recall in category I** | **Precision in category II** | **Recall in category II** |
| **Lie Detection** | 41.93% | 0.419 | 1 | - | 0 |
| **Sentiment** | 48.39 % | 0.484 | 1 | - | 0 |

Weka Output:

Table

Description automatically generated

A close-up of a document

Description automatically generated with low confidence

**Comparison**

Both the SVM model and MNB model gives better result in prediction for the sentiment analysis model. The model misclassifies more than half the test for lie detection. However, the accuracy of SVM in sentiment analysis is pretty high as compared the other three accuracies. Lie detection is a much more difficult job as it will need context and even humans tend to have a lot of difficulty in lie detection as compared to sentiment analysis.

**Gain Ratio Attribute**

The output of the top attributes for both lie detection and sentiment analysis is as follows:

Table

Description automatically generated

**Table

Description automatically generated**

It can be seen that the sentiment analysis model learned more than the lie detection model, as the ratios for the former are higher than the latter. It makes sense as words like terrible, amazing, the best, worst, etc. are highly indicative of the sentiment. But when it comes to lie detection, there are no such terms that give a lot of insight individually.