**Language Model Classifier**

**Approach:**

We built the Language Model Classifier using the unigram model and the bigram model.

**Unigram Based Language Model Classifier:**

For words that do not occur in the training data (i.e. words that occur in positive review set but not in negative review set or viceversa), we assign an <UNKNOWN> tag and calculate the probability of this tag. During classification we apply the probability of this <UNKNOWN> tag for all new/unseen words. We performed our tests both with and without Laplace Smoothing. The average success rate across all folds increased by 26% when we used Laplace smoothing.

**Results:**

|  |  |  |
| --- | --- | --- |
| **Fold** | **Success Rate** | |
| **With Smoothing** | **Without Smoothing** |
| First Fold | 81.0 | 54.75 |
| Second Fold | 80.5 | 54.25 |
| Third Fold | 78.75 | 54.0 |
| Fourth Fold | 82.25 | 53.25 |
| Fifth Fold | 79.5 | 53.75 |
| **Average of All folds** | **80.4** | **54.0** |

**Bigram Based Language Model Classifier:**

In the bigram model, we ignore any new bigrams that occur in the testing set. The average success rate across all folds increased by 23% when we used Laplace smoothing for the bigram based classifier. We have tabulated the results for the Bigram language model classifier below.

**Results:**

|  |  |  |
| --- | --- | --- |
| **Fold** | **Success Rate** | |
| **With Smoothing** | **Without Smoothing** |
| First Fold | 69.75 | 50.0 |
| Second Fold | 74.75 | 50.0 |
| Third Fold | 74.0 | 50.0 |
| Fourth Fold | 74.75 | 50.0 |
| Fifth Fold | 73.25 | 50.0 |
| **Average of All folds** | **73.3** | **50.0** |

The unigram based classifier performs better than the bigram classifier since the model becomes more constrained and specific when we use bigrams for training.

**Linear Support Vector Machine Classifier:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Trial No.** | **Type of Feature Set** | **Parameters** | | **Total Success Rate** |
| **Presence/Frequency** | **Regularization** | **Cost** |
| First Fold |  | 74.5 | 87.5 | 81.0 |
| Second Fold |  | 74.0 | 87.0 | 80.5 |
| Third Fold |  | 70.0 | 87.5 | 78.75 |
| Fourth Fold |  | 78.0 | 86.5 | 82.25 |
| Fifth Fold |  | 71.5 | 87.5 | 79.5 |
| **Average of All folds** |  | **73.6** | **87.2** | **80.4** |

We used the libsvm library for training and prediction using SVM. We trained the model using unigrams and the frequency/presence of these unigrams in the training set. We experimented with different C values as tabulated below.

**Perceptron Classifier:**

**We implemented the Perceptron Classifier based on unigrams and bigrams. We tested with different number of iterations and observed that there was no much change in the accuracy of prediction. The features used here are: presence of unigrams/bigrams and frequency of unigrams/bigrams features.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Trial No.** | **Number of Iterations** | **Type of Feature Set** | | **Total Success Rate** |
| **Presence/Frequency** | **Unigram/Bigram** |
| 1 | 10 | Presence | Unigram | 81.0 |
| 2 |  | 74.0 | 87.0 | 80.5 |
| Third Fold |  | 70.0 | 87.5 | 78.75 |
| Fourth Fold |  | 78.0 | 86.5 | 82.25 |
| Fifth Fold |  | 71.5 | 87.5 | 79.5 |
| **Average of All folds** |  | **73.6** | **87.2** | **80.4** |