**Sentiment Classification of Movie Reviews**

**Introduction**

With 1000 negative reviews and 1000 positive reviews, one can classify whether a new review is positive or negative using machine learning approaches, such as KNN, LDA, etc.

**Dataset**

The dataset I am using is IMDB Review Dataset, including 1000 positive reviews and 1000 negative reviews.

**Data preprocessing**

I removed punctuations, numbers, stop words, white spaces, and words with useless information, like “film”, “movie”, “one”, etc.

All the words are stemmed to their basic form.

**Data discovery:**

**The following are ten words with the most frequency in all the reviews.**

|  |  |
| --- | --- |
| word | frequency |
| movi | 6857 |
| like | 3998 |
| make | 3152 |
| just | 2903 |
| scene | 2638 |
| even | 2606 |
| good | 2385 |
| play | 2361 |
| stori | 2323 |
| will | 2304 |

**Plot the words with frequency more than 1850.**



**Use word clouds for visual analytics.**



**Feature engineering**

Remove words that less frequently appear in all reviews.

**Apply algorithms**

Apply KNN, NB, Linear Discriminant Analysis and SVM by 10-fold cross validation to the data set.

**Evaluation:**

**Compute the average accuracy of each algorithm.**

****

KNN:

Mean accuracy: 0.623, variation of accuracy: 0.02616189

NB:

Mean accuracy: 0.6615, variation of accuracy: 0.0297256

Linear Discriminant Analysis:

Mean accuracy: 0.748, variation of accuracy: 0.03137586

SVM:

Mean accuracy: 0.746, variation of accuracy: 0.03462497

AS show above.

Therefore, for this data set, Linear Discriminant Analysis has the highest accuracy but it has a higher variation. And we are not sure whether there is an over-fitting in LDA model.

**Use t.test to evaluate the difference between models based on different algorithms.**

If we set 0.05 as significance threshold, then the four models are different from each other. Because their p-values are less than 0.05.

**Table P\_value**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | KNN | NB | LDA | SVM |
| KNN | 1 |  |  |  |
| NB | 0.01255 | 1 |  |  |
| LDA | 1.98e-06 | 7.351e-05 | 1 |  |
| SVM | 5.897e-12 | 3.34e-14 | 1.168e-13 | 1 |

**Draw confusion matrix for each algorithm.**

KNN:

|  |  |  |
| --- | --- | --- |
| Actual attitude\prediction | 0 | 1 |
| 0 | 849 | 151 |
| 1 | 413 | 587 |

NB:

|  |  |  |
| --- | --- | --- |
| Actual attitude\prediction | 0 | 1 |
| 0 | 823 | 177 |
| 1 | 458 | 542 |

LDA:

|  |  |  |
| --- | --- | --- |
| Actual attitude\prediction | 0 | 1 |
| 0 | 815 | 185 |
| 1 | 192 | 808 |

SVM:

|  |  |  |
| --- | --- | --- |
| Actual attitude\prediction | 0 | 1 |
| 0 | 836 | 164 |
| 1 | 184 | 816 |

**Plot ROC for each algorithm and compute AUC:**

KNN: AUC = 0.7075

****

NB: AUC = 0.6825

****

LDA: AUC = 0.8115

****

SVM: AUC = 0.826

****

Therefore, by the comparison of AUC, we can conclude that SVM can better classify positive reviews and negative reviews.