**IST – 736 Text Mining**

**Multinomial naive Bayes for Fake Review Detection**

Classification of restaurant reviews by sentiment (Positive or Negative) and authenticity (True or Fake). Initial set to analysis includes Vectorization. I have used both TF-IDF and Count vectorization in both the analysis and compared the top features for each category.

On Sentiment analysis, the negative and positive words are sorted by their conditional probabilities by using Count Vectorizer and TF-IDF Vectorizer. However, it is observed that by using TF-IDF Vectorization, all the words are categorized properly for the negative category. Hence, I choose TF\_IDF vectorization for my sentiment analysis.

Top 10 Positive and Negative features for sentiment analysis: (Using TF-IDF Vectorization)

Positive Features:

**Negative Feature ranks using tfidf Vectorizer**

[(-5.227846168107487, 'asked'), (-5.222577154051034, 'terrible'), (-5.2103084299962354, 'salad'), (-5.187166247253829, 'dishes'), (-5.143151384891839, 'bad'), (-5.133996987257023, 'minutes'), (-5.1030489617666035, 'went'), (-4.938763368455484, 'place'), (-4.8430324276305585, 'food'), (-4.734849461080151, 'restaurant')]

**Positive Feature ranks using tfidf Vectorizer**

[(-5.219359097661734, 'nice'), (-5.181737794729136, 'really'), (-5.177689513221689, 'friendly'), (-5.141203834989176, 'fresh'), (-5.0652075388691316, 'amazing'), (-4.991196381116305, 'good'), (-4.9816342151921384, 'great'), (-4.842458587375431, 'restaurant'), (-4.7986462614697345, 'food'), (-4.741718621185967, 'best')]

For Authenticity analysis, it is observed that the though the importance of word is changing for both count and TF\_IDF Vectorization, both true and fake reviews have some common words. From, the analysis I could see TF-IDF vectorizer has assigned the rank for the features in a better way. Hence, I choose TF\_IDF vectorizer for this analysis as well.

**Top 10 True and Fake features for authenticity analysis: (Using TF-IDF Vectorization)**

**Fake Feature ranks using tfidf Vectorizer**

[(-5.222621534700683, 'menu'), (-5.204141206095463, 'minutes'), (-5.194766449563898, 'want'), (-5.191439930725854, 'like'), (-5.169589079803627, 'experience'), (-5.091523346372281, 'service'), (-5.036297924807053, 'place'), (-4.9902384138165345, 'best'), (-4.931710867417781, 'restaurant'), (-4.8395399172187705, 'food')]

**True Feature ranks using tfidf Vectorizer**

[(-5.3097769800584675, 'went'), (-5.279450072677803, 'bad'), (-5.2643763095714355, 'dishes'), (-5.242117858118911, 'really'), (-5.22604674402891, 'salad'), (-5.209498165662844, 'ordered'), (-5.123975695678814, 'place'), (-5.008451736461623, 'good'), (-4.801344505308265, 'food'), (-4.653277927142044, 'restaurant')]

Model Performance for true fake detection using TF-IDF vectorization:

|  |  |  |  |
| --- | --- | --- | --- |
| **Authenticity** | **Precision** | **Recall** | **F1-Score** |
| f | 0.50 | 0.77 | 0.61 |
| t | 0.62 | 0.33 | 0.43 |
| Accuracy – 0.54 |  |  |  |

Model Performance for sentiment analysis using TF-IDF vectorization:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sentiment** | **Precision** | **Recall** | **F1-Score** |
| n | 1.00 | 0.73 | 0.85 |
| p | 0.76 | 1.00 | 0.87 |
| Accuracy – 0.86 |  |  |  |

From both the tables, it is observed that Multinomial Naïve Bayes algorithm, classified the Positive and Negative reviews accurately. Whereas it was not able to classify the fake and true reviews accurately.

**Error Analysis:**

For sentiment analysis, there are **4** errors in predicting the Negative review as positive for TFIDF vectorizer, whereas no positive reviews are predicted as negative. On analysis, it is observed that there are certain positive words like “great”,” comfort” in the negatively labelled reviews which might have categorized them into Positive category.

For authenticity analysis, there are **3** errors that are predicted as true for fake. Also, there are **3** errors in predicting fake as true. There can be errors when most frequent true words are part of fake review or vice versa. Cross validation **with 5 folds** is also used in the analysis, which resulted in **88 percent** accuracy for Sentiment analysis and **58 percent** for authenticity analysis

On comparing, Sentiment analysis and authenticity analysis for restaurant review, it is observed that the positive and negative reviews are classified with high accuracy with very less false positives and false negatives. Whereas, for authenticity analysis, the classification for true and fake reviews is less accurate, with only 58 percent accuracy with high false negatives and false positives.

Classification of fake and true reviews is more difficult compared to the sentiment analysis. Because both the categories (true and fake) have common words. Hence, the algorithm was not able to distinguish between them.