**IST 736**

**Assignment 2**

**Name: Yashaswini Kulkarni**

**SUID: 933791832**

**Executive Summary**

The assignment aims at conducting a comprehensive comparison of the performance of the MNB (Multinomial Naïve Bayes) machine learning algorithm and the SVM (Support Vector Machine) machine learning algorithm for two tasks i.e., sentiment analysis and authenticity classification. These tasks are performed by using multiple vectorizers specifically TF-IDF vectorizer, count vectorizer and binary/bigram vectorizer.

For this, the first step is preparing the data. In text pre-processing, the data underwent a series of essential steps to enhance its suitability for analysis. Non-alphabetic characters were systematically removed, and all text was uniformly converted to lowercase to ensure consistency. The process then involved tokenization, breaking down the text into individual units, followed by lemmatization to reduce words to their base forms. These steps collectively aimed to refine the textual data, making it more manageable and conducive for subsequent analysis.

The Multinomial Naive Bayes (MNB) model, a probabilistic classification algorithm, was implemented to predict sentiments and assess authenticity. It leverages the probabilities of different features in the dataset to make predictions, making it particularly effective for text classification tasks. The model was trained and tested using a dataset, and three distinct vectorization techniques—count, TF-IDF, and binary—were utilized for sentiment analysis. For authenticity analysis, count, TF-IDF, and bigram vectorizers were employed. The model's performance was rigorously assessed through a 5-fold cross-validation, evaluating metrics such as precision, recall, and others to ensure a robust understanding of its predictive capabilities.

The Support Vector Machine (SVM) model, a powerful classification algorithm, was implemented for sentiment and authenticity analysis. Operating on the principles of margin maximization, SVM effectively separates classes in high-dimensional spaces. The model underwent training and testing on a dataset, employing three distinct vectorization techniques—count, TF-IDF, and binary—for sentiment analysis. In the context of authenticity analysis, count, TF-IDF, and bigram vectorizers were utilized. The model's performance underwent rigorous evaluation through a 5-fold cross-validation process, comprehensively assessing precision, recall, and other metrics to ensure a robust evaluation of its predictive capabilities in both sentiment and authenticity tasks.

For **sentiment analysis**, the **Multinomial Naïve bayes** machine learning algorithm gives the **best** results. Additionally, using the **TF-IDF vectorizer** gives the most desired results with this algorithm. However, for the **authenticity analysis**, **both** these models perform **poorly**. They gave very bad accuracies and a large number of incorrect predictions. From the feature analysis, it is clear

**Model Evaluation**

Model 1: MNB Model with Tf-Idf vector for sentiment analysis

Accuracy: 0.9285714285714286

|  |
| --- |
| precision recall f1-score support |
|  |
| n 1.00 0.88 0.93 16 |
| p 0.86 1.00 0.92 12 |

Model 2: MNB Model with the count vectorizer for sentiment analysis

Accuracy: 0.8571428571428571

|  |
| --- |
| precision recall f1-score support |
|  |
| n 1.00 0.75 0.86 16 |
| p 0.75 1.00 0.86 12 |

Model 3: MNB Model using Binary vectorizer for sentiment analysis

Accuracy: 0.8214285714285714

|  |
| --- |
| precision recall f1-score support |
|  |
| n 0.92 0.75 0.83 16 |
| p 0.73 0.92 0.81 12 |

Model 4: MNB Model using Tf-Idf vector for authenticity analysis

Accuracy: 0.5357142857142857

|  |
| --- |
| precision recall f1-score support |
|  |
| f 0.56 0.60 0.58 15 |
| t 0.50 0.46 0.48 13 |

Model 5: MNB Model using Count Vectorizer for authenticity analysis

Accuracy: 0.5714285714285714

|  |
| --- |
| precision recall f1-score support |
|  |
| f 0.64 0.47 0.54 15 |
| t 0.53 0.69 0.60 13 |

Model 6: MNB Model with Bigram vectorizer for authenticity Analysis

Accuracy (MNB with Bigram Vectorizer): 0.42857142857142855

|  |
| --- |
| precision recall f1-score support |
|  |
| f 0.47 0.53 0.50 15 |
| t 0.36 0.31 0.33 13 |

Model 7: SVM Model for sentiment analysis using tf-idf model

Accuracy: 0.8928571428571429

|  |
| --- |
| precision recall f1-score support |
|  |
| n 0.93 0.88 0.90 16 |
| p 0.85 0.92 0.88 12 |

Model 8: SVM Model for sentiment analysis using count vectorizer

Accuracy (Linear Kernel, Count Vectorizer): 0.8571428571428571

|  |
| --- |
| precision recall f1-score support |
|  |
| n 0.88 0.88 0.88 16 |
| p 0.83 0.83 0.83 12 |

Model 9: SVM Model for sentiment analysis using Bigram vectorizer

Accuracy (SVM with Binary Vectorizer): 0.7857142857142857

|  |
| --- |
| precision recall f1-score support |
|  |
| n 0.86 0.75 0.80 16 |
| p 0.71 0.83 0.77 12 |

Model 10: SVM Model for authenticity analysis using Tf-Idf vectorizer

Accuracy: 0.5714285714285714

|  |
| --- |
| precision recall f1-score support |
|  |
| f 0.62 0.53 0.57 15 |
| t 0.53 0.62 0.57 13 |

Model 11: SVM Model for authenticity analysis using the Count vectorizer

Accuracy (Linear Kernel, Count Vectorizer): 0.5

|  |
| --- |
| precision recall f1-score support |
|  |
| f 0.55 0.40 0.46 15 |
| t 0.47 0.62 0.53 13 |

Model 12: SVM Model for authenticity analysis using the bigram vectorizer

Accuracy (Bigram Vectorizer): 0.42857142857142855

|  |
| --- |
| precision recall f1-score support |
|  |
| f 0.40 0.13 0.20 15 |
| t 0.43 0.77 0.56 13 |

In summary: (accuracies in percentages)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Tf-Idf Vectorizer** | | **Count Vectorizer** | | **Binary/Bigram Vectorizer** | |
|  | **MNB** | **SVM** | **MNB** | **SVM** | **MNB** | **SVM** |
| **Sentiment Analysis** | 92.8 | 89.3 | 85.7 | 85.7 | 82.1 | 42.9 |
| **Authenticity Analysis** | 53.6 | 57.1 | 57.1 | 50 | 42.9 | 42.9 |

**Feature Analysis**

Feature analysis plays a pivotal role in understanding the discriminative elements that classifiers leverage to make predictions. In the context of sentiment and authenticity analysis, examining the top features identified by Support Vector Machines (SVM) and Multinomial Naive Bayes (MNB) sheds light on the distinctive characteristics they prioritize. SVM, known for its effectiveness in capturing complex relationships, identifies features like "restaurant have" and "best restaurant," emphasizing the importance of positive sentiments and high-quality establishments. On the other hand, MNB, relying on probabilistic principles, highlights features such as "the food" and "the place," suggesting a focus on more specific aspects of reviews. The differences in these top features underscore the unique approaches these algorithms take, showcasing the nuanced preferences in sentiment and authenticity classification.

**Error Analysis**

The error analysis revealed nuanced patterns in the model predictions. Instances of misclassification in sentiment analysis often stemmed from the model's struggle with subtle contextual cues, leading to occasional misinterpretation of negative sentiments as positive or vice versa. In authenticity analysis, the misclassification instances highlighted challenges in distinguishing between genuine and fabricated content, particularly when narratives contained mixed sentiments. This analysis underscores the importance of continuous refinement to address the intricacies of language and context, contributing to the ongoing enhancement of model performance in both sentiment and authenticity tasks.

**Sentiment Analysis vs Fake review Classification**

Comparatively, authenticity classification is far more difficult than sentiment analysis. Sentiment classification encounters challenges in discerning emotional nuances, leading to occasional misinterpretations of positive or negative sentiments. On the other hand, fake review classification proves more demanding due to the inherent complexity of distinguishing between genuine and fabricated content. Factors such as limited diversity in training data, the subjective nature of authenticity, and the necessity for more tailored feature representations contribute to lower accuracies in fake review classification. This highlights that while sentiment analysis involves capturing emotional tones, fake review classification requires a deeper understanding of authenticity and calls for nuanced approaches to address distinct intricacies, leading to differences in the difficulty levels of the two tasks.