#### **ECSE 551 Machine Learning for Engineers**

Mini-Project 3

Lab Report

Fei Peng

260712440

Yukai Zhang

260710915

Yi Zhu

260716006

**Link to LaTex:** <https://www.overleaf.com/8783635833pxrrcnxbrgvf> (updated)

1. Abstract

2. Introduction

3. Dataset

4. Proposed Approach:

5. Results:

6. Discussion and conclusion:

7. Statement of Contributions

Reference

Appendix

Appendix A: Results

Table A.1: Performance of Bernoulli Naïve Bayes implemented by ourselves.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Vectorizer | Normalization | Max Features | Time Spent (second) | Accuracy (%) |
| CountVectorizer | Yes | Unlimited | 1804.497 | 74.288 |
| CountVectorizer with Stopwords | Yes | 5000 | 168.842 | 81.316 |
| CountVectorizer with Stopwords | Yes | Unlimited | 1285.886 | 78.890 |
| CountVectorizer with Stemming | Yes | Unlimited | 1049.033 | 72.224 |
| CountVectorizer with Lemmatization | Yes | Unlimited | 1750.286 | 72.906 |

Table A.2: Performance of Logistic Regression.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Vectorizer | Normalization | Max Features | Time Spent (second) | Accuracy (%) |
| CountVectorizer | Yes | Unlimited | 64.152 | 82.214 |
| CountVectorizer | No | Unlimited | 117.682 | 86.695 |
| CountVectorizer with Stopwords | Yes | Unlimited | 47.928 | 86.609 |
| CountVectorizer with Stopwords | No | Unlimited | 37.534 | 87.455 |
| TF-IDF Vectorizer | Yes | Unlimited | 116.797 | 89.613 |
| TF-IDF Vectorizer | No | Unlimited | 116.167 | 89.233 |
| CountVectorizer with Stemming | Yes | Unlimited | 189.210 | 81.748 |
| CountVectorizer with Stemming | No | Unlimited | 236.312 | 85.970 |
| CountVectorizer with Lemmatization | Yes | Unlimited | 743.057 | 81.471 |
| CountVectorizer with Lemmatization | No | Unlimited | 809.459 | 86.039 |

Table A.3: Accuracy of Multinomial Naïve Bayes.

|  |  |  |  |
| --- | --- | --- | --- |
| Vectorizer | Normalization | Max Features | Accuracy (%) |
| CountVectorizer | Yes | Unlimited | 58.012 |
| CountVectorizer with Stopwords | Yes | Unlimited | 73.441 |
| TF-IDF Vectorizer | Yes | Unlimited | 70.100 |
| CountVectorizer with Stemming | Yes | Unlimited | 58.962 |
| CountVectorizer with Lemmatization | Yes | Unlimited | 57.503 |

Table A.4: Accuracy of Support Vector Machine.

|  |  |  |  |
| --- | --- | --- | --- |
| Vectorizer | Normalization | Max Features | Accuracy (%) |
| CountVectorizer | Yes | Unlimited | 84.234 |
| CountVectorizer with Stopwords | Yes | Unlimited | 87.351 |
| TF-IDF Vectorizer | Yes | Unlimited | 88.715 |
| CountVectorizer with Stemming | Yes | Unlimited | 83.543 |
| CountVectorizer with Lemmatization | Yes | Unlimited | 83.284 |

Table A.5: Accuracy of Random Forest.

|  |  |  |  |
| --- | --- | --- | --- |
| Vectorizer | Normalization | Max Features | Accuracy (%) |
| CountVectorizer | Yes | Unlimited | 30.358 |
| CountVectorizer with Stopwords | Yes | Unlimited | 32.283 |
| TF-IDF Vectorizer | Yes | Unlimited | 31.842 |
| CountVectorizer with Stemming | Yes | Unlimited | 38.111 |
| CountVectorizer with Lemmatization | Yes | Unlimited | 34.528 |

Table A.6: Accuracy of Decision Tree.

|  |  |  |  |
| --- | --- | --- | --- |
| Vectorizer | Normalization | Max Features | Accuracy (%) |
| CountVectorizer | Yes | Unlimited | 67.087 |
| CountVectorizer with Stopwords | Yes | Unlimited | 70.471 |
| TF-IDF Vectorizer | Yes | Unlimited | 66.906 |
| CountVectorizer with Stemming | Yes | Unlimited | 66.517 |
| CountVectorizer with Lemmatization | Yes | Unlimited | 65.153 |

Table A.7: Accuracy of AdaBoost.

|  |  |  |  |
| --- | --- | --- | --- |
| Vectorizer | Normalization | Max Features | Accuracy (%) |
| CountVectorizer | Yes | Unlimited | 68.434 |
| CountVectorizer with Stopwords | Yes | Unlimited | 67.968 |
| TF-IDF Vectorizer | Yes | Unlimited | 68.546 |
| CountVectorizer with Stemming | Yes | Unlimited | 70.868 |
| CountVectorizer with Lemmatization | Yes | Unlimited | 69.807 |

Table A.8: Accuracy of K-Neighbors.

|  |  |  |  |
| --- | --- | --- | --- |
| Vectorizer | Normalization | Max Features | Accuracy (%) |
| CountVectorizer | Yes | Unlimited | 46.194 |
| CountVectorizer with Stopwords | Yes | Unlimited | 65.447 |
| TF-IDF Vectorizer | Yes | Unlimited | 79.019 |
| CountVectorizer with Stemming | Yes | Unlimited | 48.860 |
| CountVectorizer with Lemmatization | Yes | Unlimited | 47.643 |

Table A.9: Accuracy of Neural Network.

|  |  |  |  |
| --- | --- | --- | --- |
| Vectorizer | Normalization | Max Features | Accuracy (%) |
| CountVectorizer | Yes | Unlimited | 89.803 |
| CountVectorizer with Stopwords | Yes | Unlimited | 89.898 |
| TF-IDF Vectorizer | Yes | Unlimited | 90.494 |
| CountVectorizer with Stemming | Yes | Unlimited | 88.361 |
| CountVectorizer with Lemmatization | Yes | Unlimited | 88.517 |

Table A.10: Performance of Bernoulli Naïve Bayes implemented by Sklearn.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Vectorizer | Normalization | Max Features | Time Spent (second) | Accuracy (%) |
| CountVectorizer | Yes | Unlimited | 5.479 | 73.934 |
| CountVectorizer with Stopwords | Yes | Unlimited | 4.885 | 74.633 |
| CountVectorizer with Stopwords | Yes | 5000 | 4.717 | 83.448 |
| TF-IDF Vectorizer | Yes | Unlimited | 5.328 | 73.891 |
| CountVectorizer with Stemming | Yes | Unlimited | 139.701 | 78.216 |
| CountVectorizer with Lemmatization | Yes | Unlimited | 689.932 | 76.826 |

Table A.11: Time Required for Classifiers with TF-IDF Vectorizer

|  |  |
| --- | --- |
| Classifier | Time Spent (second) |
| Multinomial NB | 5.411 |
| Support Vector Machine | 321.477 |
| Random Forest | 7.130 |
| Decision Tree | 33.605 |
| Ada Boost | 70.709 |
| K-Neighbors | 12.815 |
| Neural Network | 1379.423 |

Table A.12: Performance of Ensemble Classifier.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier | Base Learner 1 | Base Learner 2 | Base Learner 3 | Time Spent (s) | Accuracy (%) |
| Voting Classifier | Neural Classifier | Logistic Reg. | SVM | 3707.132 | 90.511 |
| Stacking Classifier | Neural Classifier | SVM | Logistic Reg. | 19795.276 | 91.029 |

Appendix B: Code