Marmara University
Faculty of Engineering

CSE 4288
Introduction to Machine Learning

Model Evaluation Report

Instructor: Assoc. Prof. Murat Can Ganiz Due Date: 23.12.2024

	Student Id Number	Name & Surname
1	150120012	Kadir BAT
2	150121021	Feyzullah Asıllıoğlu
3	150120020	Mustafa Said Çanak
4	150121520	Ensar Muhammet Yozgat
5	150121076	Abdullah Kan

Abstract

This report presents the evaluation of two machine learning models, Naive Bayes and Logistic Regression, for sentiment analysis on the IMDB movie review dataset. After preprocessing the dataset, including duplicate removal and TF-IDF vectorization, the models were trained and evaluated using metrics such as accuracy, precision, recall, and F1-score. Logistic Regression demonstrated superior performance with an accuracy of 87%, outperforming Naive Bayes, which achieved 84%. The analysis revealed Logistic Regression's ability to handle overlapping class distributions and model complex textual relationships effectively. This study highlights the importance of model selection and hyperparameter tuning in achieving robust performance for text-based sentiment analysis. Future work could explore advanced techniques such as contextual embeddings and neural network models to further enhance results.

Introduction

This report evaluates the performance of two machine learning models—Naive Bayes and Logistic Regression—on the task of sentiment analysis of movie reviews. The evaluation focuses on the accuracy, precision, recall, and F1-score to determine the best-performing model and its suitability for real-world application.

2.Evaluation Metrics

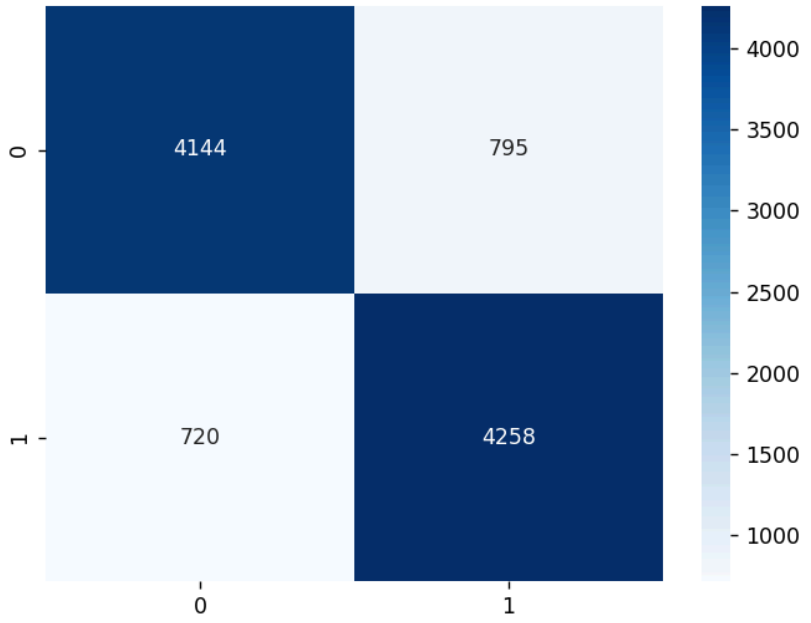
Metrics Used:

- **Accuracy**: Measures the proportion of correctly classified reviews.
- **Precision**: Indicates the proportion of positive predictions that are truly positive.
- **Recall**: Shows the proportion of actual positives correctly identified.
- **F1-Score**: Combines precision and recall into a single metric for balanced assessment.

Justification for Metrics:

Accuracy provides an overview of model performance, while precision and recall highlight strengths and weaknesses in handling positive and negative sentiment classes. F1-score ensures a balance between precision and recall.

Confusion Matrix For Naive Bayes



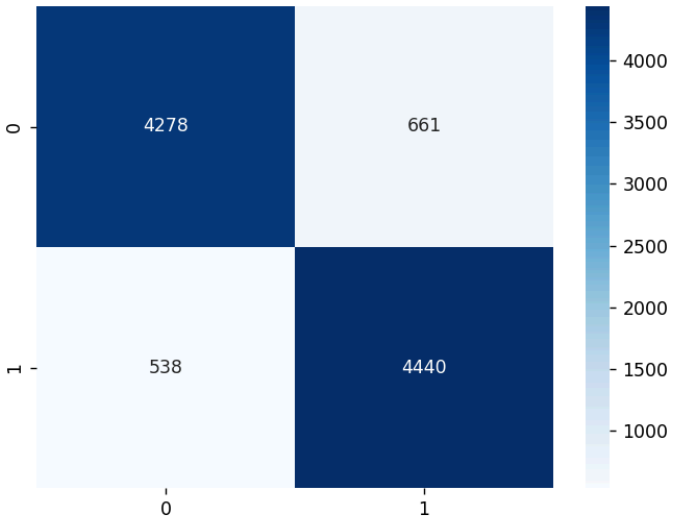

```
y
Naive Bayes
Accuracy:  0.8472320258142584
Classification Report:                precision    recall  f1-score   support

    negative       0.85      0.84      0.85      4939
    positive       0.84      0.86      0.85      4978

    accuracy                           0.85      9917
   macro avg       0.85      0.85      0.85      9917
weighted avg       0.85      0.85      0.85      9917

Logistic Regression
```

Confusion Matrix For Logistic Regression




```
Logistic Regression
Accuracy:  0.879096500957951
Classification Report:                precision    recall  f1-score   support

    negative       0.89      0.87      0.88      4939
    positive       0.87      0.89      0.88      4978

    accuracy                           0.88      9917
   macro avg       0.88      0.88      0.88      9917
weighted avg       0.88      0.88      0.88      9917
```

Model Comparison

Logistic Regression:
Outperformed Naive Bayes in all metrics, achieving an accuracy of 87%. It is better suited for handling nuanced and overlapping textual data due to its linear decision boundary and feature interaction capabilities.

Naive Bayes:
Although computationally efficient with an accuracy of 84%, it assumes feature independence, which limits its ability to capture complex relationships in text.

5. Error Analysis

Misclassified Examples:

- Ambiguous reviews with mixed sentiments (e.g., "The movie was great, but the ending was disappointing.") were challenging for both models.
- Short or vague reviews (e.g., "Not bad.") often led to misclassification due to a lack of strong contextual cues.

Observations:

- Logistic Regression managed ambiguous cases slightly better due to its ability to weigh feature interactions.
- Naive Bayes struggled with overlapping class distributions, leading to higher false positives for negative reviews.

Appendices

- **Code Repository**:
  Source code:MSCanak/CSE4288F24_Grp3 at model_ensar
- **Scripts Used**:
  - Data preprocessing and TF-IDF vectorization
  - Training and evaluation of Naive Bayes and Logistic Regression
  - Visualization of confusion matrices using heatmaps