SE 4475 Final Report

Seden Canpolat 20070001044

The goal of this project was to classify text documents which are Turkish-language articles based on their authors.

The use of the Zemberek library was preferred because of the use of a dataset containing Turkish words. Tokenization and stemming were applied using a Turkish language tokenizer and morphological analyzer.

Logistic Regression is used as a model in this assignment. It’s a linear classifier that predicts the probability of different classes using the logistic function. It is used for binary classification and extended to multi-class problems.

**Parameters used in Logistic Regression:**

* **max\_ iter: 5000** => The maximum number of iterations for convergence was set to 5,000 to handle the dataset size.

The text data was converted into numerical representations using Term Frequency-Inverse Document Frequency (TF-IDF).

**Parameters used in TF-IDF Vectorization:**

* **max\_features:** **10000** => The top 10,000 terms were retained based on their TF-IDF scores.
* **ngram\_range: (1, 2)** => Both unigrams and bigrams were used for vectorization.
* **sublinear\_tf: True** =>Sublinear term frequency scaling was applied to reduce the effect of high-frequency terms

In this project, 5-Fold Cross-Validation was used as the validation strategy. The dataset was split into five folds using stratified sampling to keep the class distribution balanced. Each fold became a test set once, while the remaining four folds were used for training. The evaluation metrics were then averaged across all folds.

The model's performance was evaluated using precision, recall, and F1-score for each class during validation. The average scores across all classes were also calculated.

In conclusion, Logistic Regression with TF-IDF performed well on the dataset. 5-Fold Cross-Validation ensured a reliable performance evaluation. With all of this, the performance results were excellent. This can have a negative aspect, as the high-performance results may indicate a risk of overfitting.