WIDS-2025
Project Report

Boda Prabanjan Jadav
24B1001
Mentor : Aditya Sanapala

Contents

1 Introduction

In this assignment, I worked on sentiment analysis using Natural Language Processing techniques. The main objective was to understand how text data can be processed and classified using machine learning models. The tasks assigned in this project include implementing tokenization and stemming, building a TF-IDF based sentiment classifier, and exploring Hugging Face datasets.

—

2 Dataset and Hugging Face Exploration

I explored the Hugging Face `datasets` library to access publicly available NLP datasets. Using this library, I loaded the SST-2 (Stanford Sentiment Treebank) dataset, which is a binary sentiment classification dataset.

The SST-2 dataset contains sentences labeled as either positive or negative. I used the predefined training and validation splits provided by Hugging Face. This helped me avoid manual data splitting and ensured a standard and reliable dataset setup.

—

3 Tokenization and Stemming

As part of text preprocessing, I implemented tokenization and stemming using Python and the NLTK library. Tokenization was performed to split each sentence into individual words. This allows the text to be processed at the word level instead of as raw sentences.

After tokenization, I applied stemming using the Porter Stemmer. Stemming reduces words to their root form and helps handle different variations of the same word. These steps reduce vocabulary size and improve the efficiency of the classification model.

—

4 Feature Extraction using TF-IDF

To convert text data into numerical form, I used the TF-IDF (Term Frequency–Inverse Document Frequency) technique. TF-IDF measures how important a word is to a sentence relative to the entire dataset.

I applied TF-IDF vectorization to the input sentences so that they could be used by a machine learning model. This representation captures important words while reducing the influence of very common words.

—

5 Sentiment Classification Model

For sentiment classification, I built a text classifier using a Linear Support Vector Machine (Linear SVM). The TF-IDF vectorizer and the classifier were combined using a pipeline so that feature extraction and classification occur together.

The model was trained using the SST-2 training data and evaluated on the validation data. This approach satisfies the requirement of building a TF-IDF based sentiment classifier for the SST-2 dataset.

—

6 Evaluation

After training the model, I evaluated its performance on the validation dataset. I used standard classification metrics such as precision, recall, and F1-score to measure how well the model predicts sentiment labels.

These metrics provide a clear understanding of the strengths and limitations of the classifier.

—

7 Conclusion

In this assignment, I successfully completed all the assigned tasks. I explored the Hugging Face datasets library and loaded the SST-2 dataset. I implemented tokenization and stemming using Python and NLTK. Finally, I built a TF-IDF based sentiment classifier using a Linear SVM model and evaluated its performance.

This project helped me understand the complete workflow of text preprocessing, feature extraction, and sentiment classification using machine learning.