Team_ID: T13

Project name: Sentiment Analysis of movie reviews

Member ID	Member name
2021170261	شهد حسن اكرام صالح فريد
2021170507	مريم صلاح سالم شعبان
2021170268	ضحى عادل حسن مصطفى
2021170183	دينا طارق عبد الحميد عبد الحميد
2021170504	مريم رضا عبد القادر محمود
2021170672	مريم أحمد طه أحمد الجوهري

Sentiment Analysis with Support Vector Machines (SVM)

Introduction

This document presents a sentiment analysis task using Support Vector Machines (SVM). The goal is to classify movie reviews as positive or negative based on their content.

Dataset

The dataset consists of movie reviews collected from the NLTK corpus. Each review is labeled as either positive or negative. The dataset is preprocessed and tokenized before being used for training the SVM classifier.

Steps:

1. Data Preprocessing

- Read movie reviews from text files stored in separate directories for positive and negative reviews.
- Tokenize the reviews using whitespace tokenization.
- Clean the tokens by removing special characters.
- Remove stop words using NLTK's English stop word list.
- Lemmatize the tokens using WordNet Lemmatizer with Part-of-Speech (POS) tagging.

2. Feature Extraction

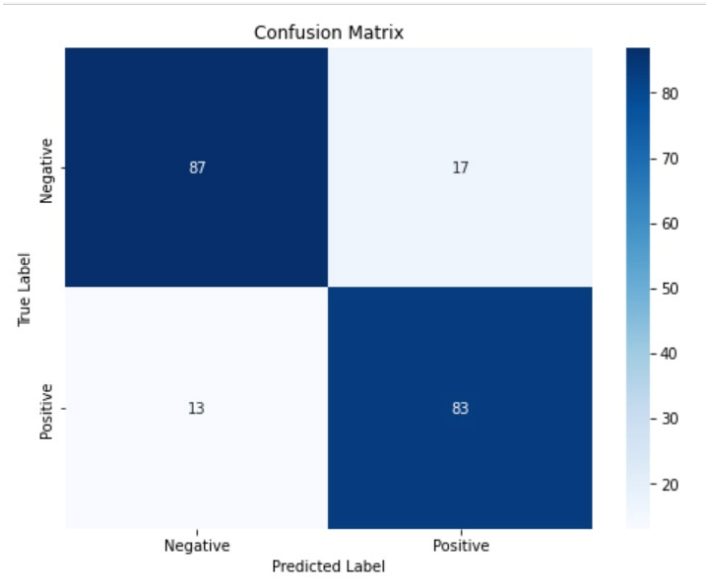
- Convert the preprocessed tokens into a list of strings.
- Initialize the TF-IDF vectorizer to convert the text data into numerical feature vectors.
- Fit and transform the documents using the TF-IDF vectorizer.

3. Model Training

- Split the dataset into training and testing sets.
- Initialize an SVM classifier with a linear kernel.
- Train the SVM classifier using the training data.

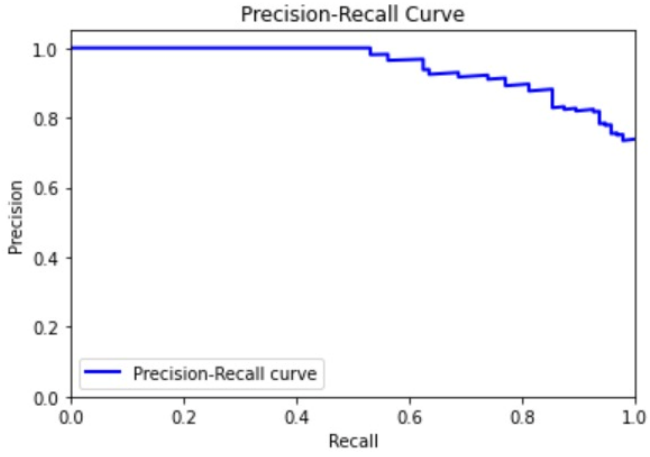
4. Model Evaluation

- Make predictions on the test data using the trained SVM classifier.
- Evaluate the model performance using accuracy score and classification report.
- Visualize the confusion matrix to analyze the model's performance.



5. Precision-Recall Curve

- Compute the precision-recall curve to further evaluate the model's performance.
- Plot the precision-recall curve for visualization.



6. Model Persistence

- Save the trained SVM model to a file using the pickle module for future use.
- Load the saved model from the file to make predictions.

Results:

- **Accuracy:** The SVM classifier achieved an accuracy of 83% on the test set.
- **Precision-Recall Curve:** The precision-recall curve indicates the trade-off between precision and recall for different thresholds.
- **Model Persistence:** The trained SVM model was successfully saved to a file and loaded back for making predictions. The loaded model's performance matches the original model.

Conclusion:

The SVM classifier trained on movie reviews achieved 83% accuracy in classifying reviews as positive or negative. The precision-recall curve provides insights into the model's performance across different thresholds. By persisting the trained model, it can be reused for making predictions on new data without retraining.