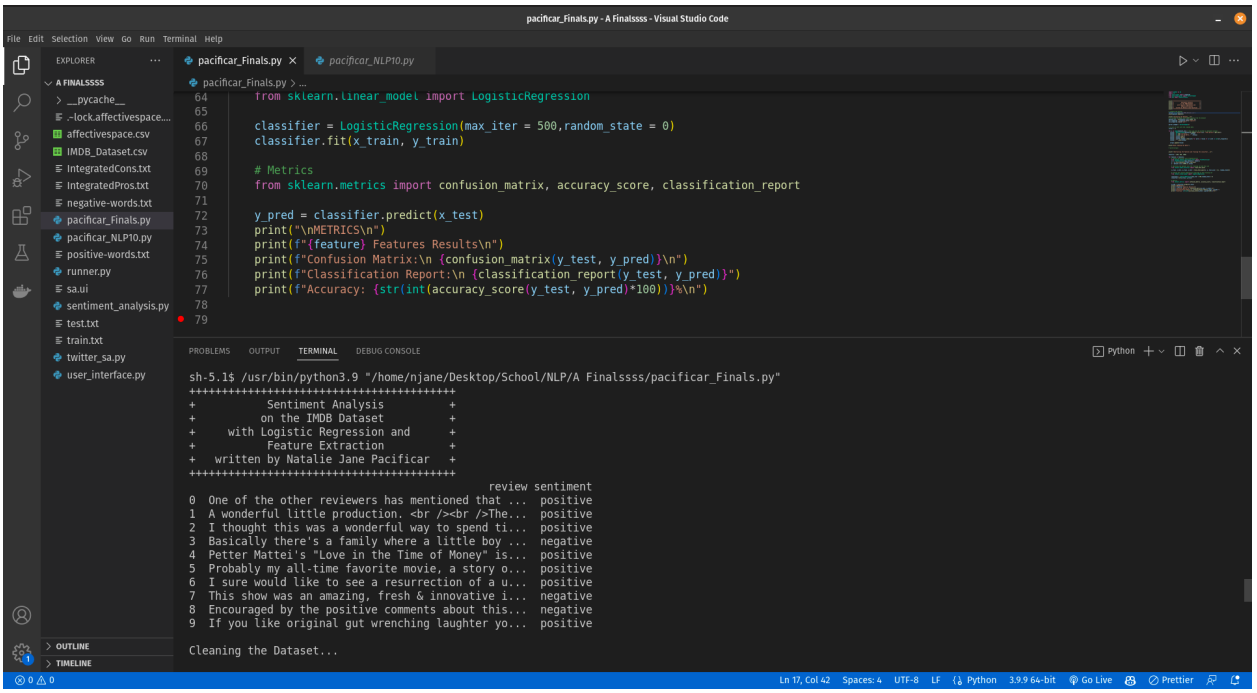
Natalie Jane Pacificar BSCS 3-B

FINALS PROJECT ON NLP

Dataset: IMDB Movie Reviews Dataset
Retrieved from:
https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews



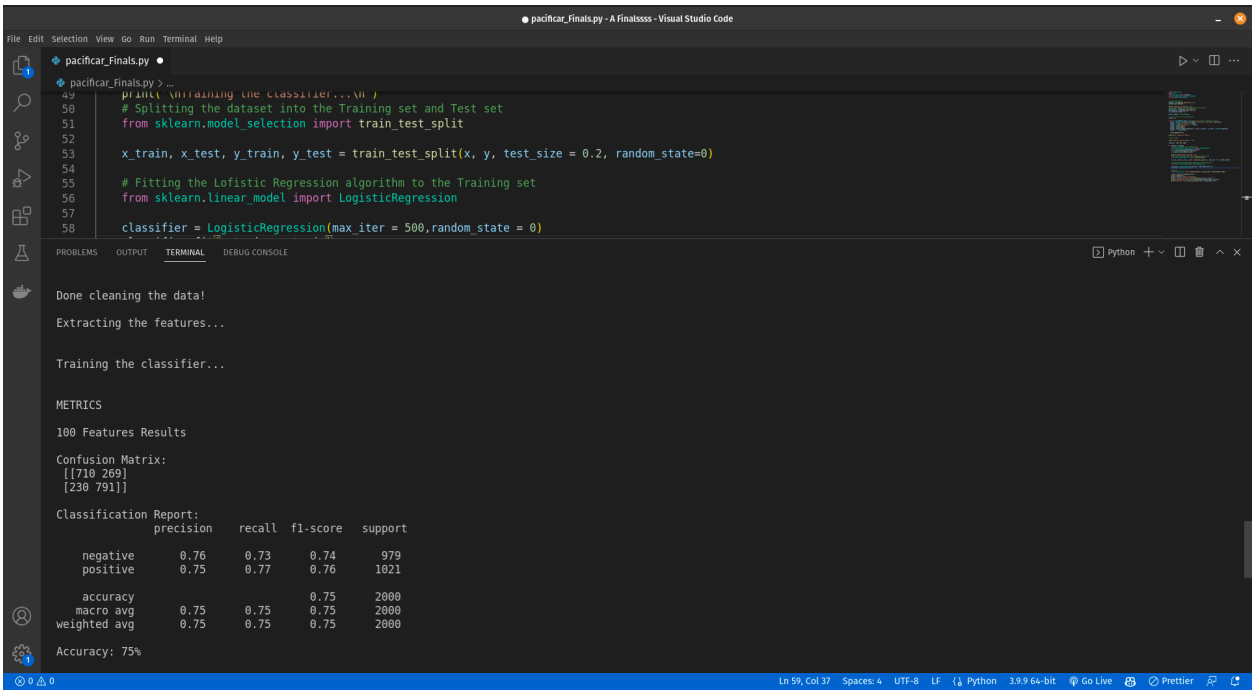
1. Cleaned the dataset by removing the stopwords, parsing the data with BeautifulSoup, using regex to remove unnecessary characters, and using the Porter Stemmer to reduce the words into their lemmas.

2. Extracted Features to increase the accuracy of our algorithm. I tried different numbers of features to see if there are significant changes in our scores.

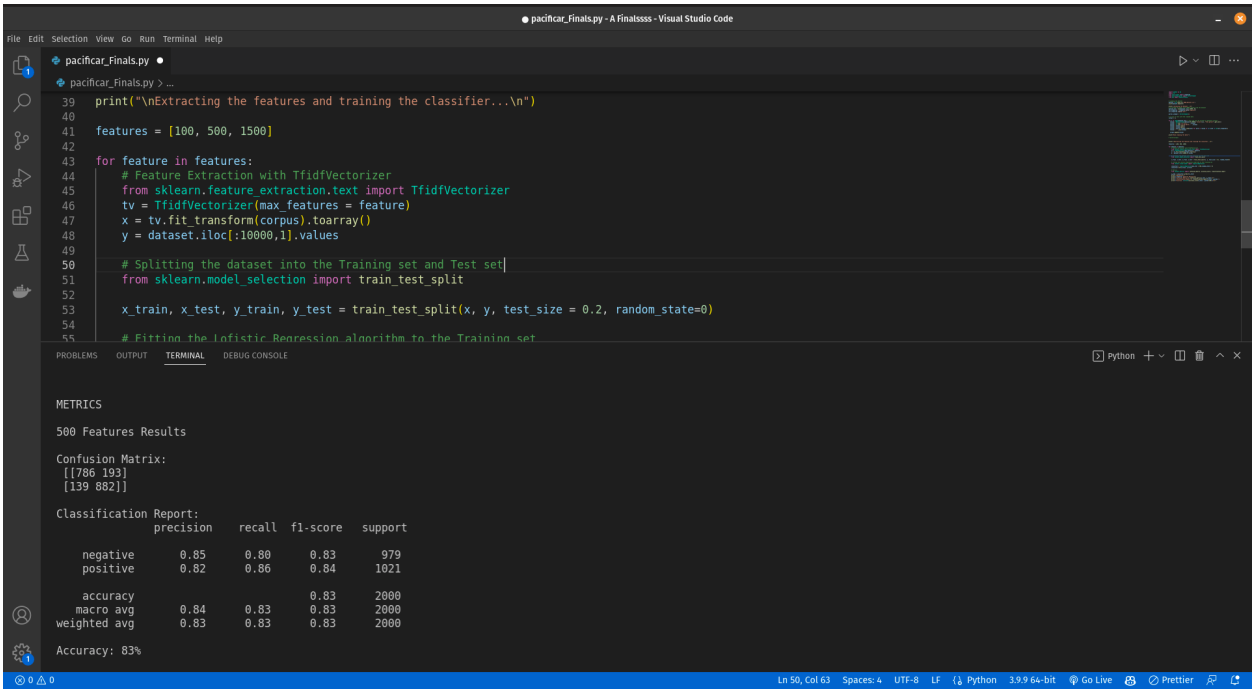
3. For the classifier, I used Logistic Regression to fit the algorithm to our training set as well as to predict the sentiment.

Natalie Jane Pacificar BSCS 3-B

Our algorithm training on **100** Features. It only achieved **75%** accuracy.

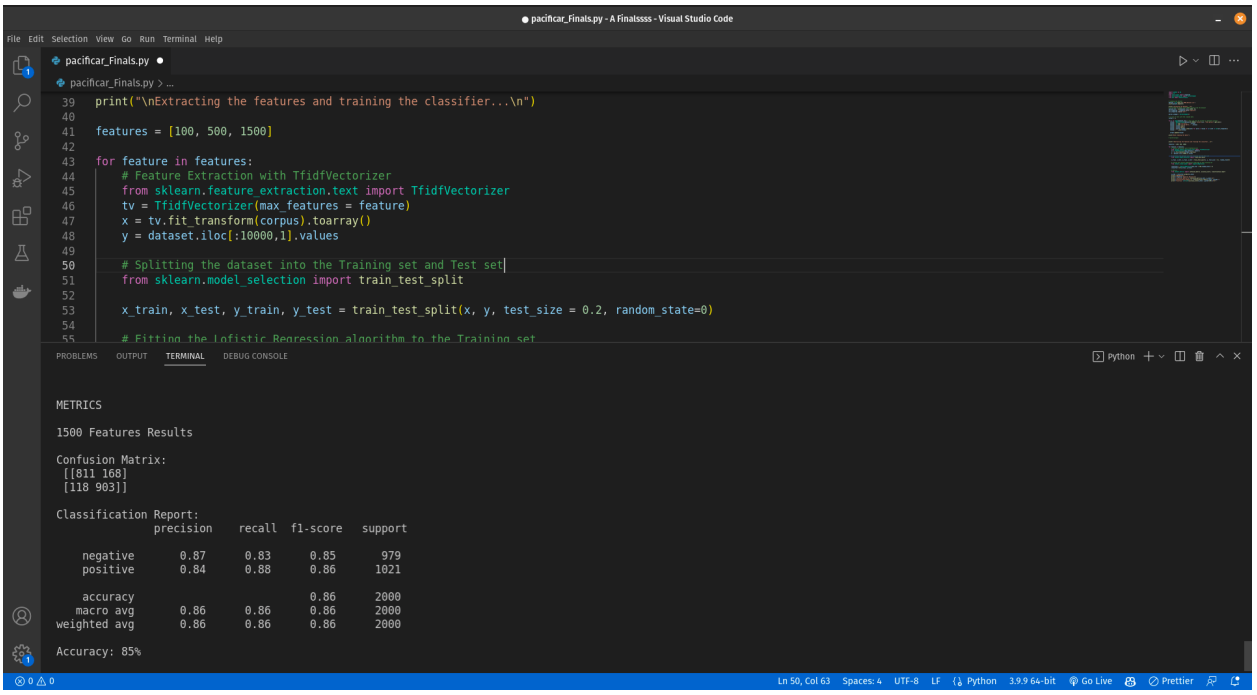


Our algorithm training on **500** Features. It achieved **83%** accuracy.

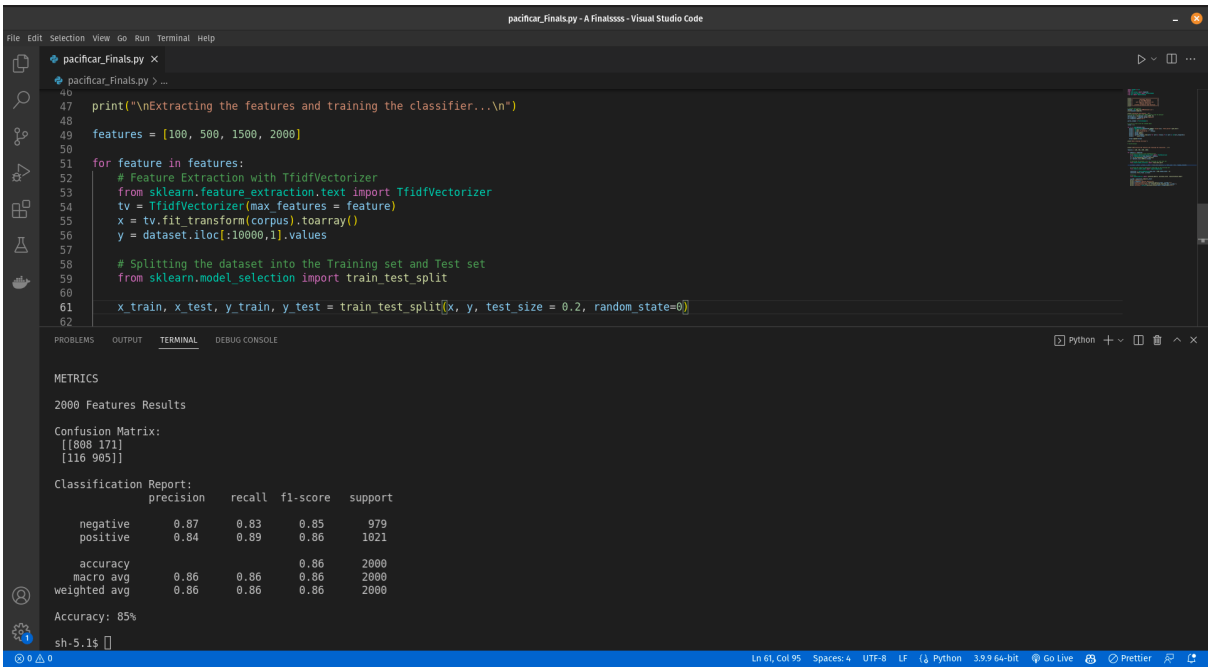


Our algorithm training on **1500** Features. It achieved **85%** accuracy.

Natalie Jane Pacificar BSCS 3-B



Our algorithm training on 2000 Features. It still achieved **85%** accuracy.



We trained our algorithm with different numbers of features. We found out that the algorithm reached its peak with **1500 features**. The number of False Positives and False Negatives was greatly reduced which means that the accuracy also increased together with the scores in the classification report as we increase the number of features up to 1500 features.