Report

Introduction:

Task: To build a sentiment analysis model to predict whether a movie review is positive or negative.
Using NLP and other ML techniques, we can achieve this task.

Problem Definition and Algorithm:

The task is to build a sentiment analysis model to predict whether the movie review is positive or negative.
For any task involving NLP, there are several factors that influence the metrics of the system like size of the text corpus, vectorization technique, and choice of model to name a few.

The first step would be to perform EDA on the dataset which involves checking for missing data, analysing the data and then splitting the data into testing and training data.
The next step would be vectorization, where the choice of TF-IDF is made. It offers the best accuracy in comparison with Bag Of Words and Bag of N-Grams.
The next step would be training the model and then finally checking for accuracy.

Results:

```
Training accuracy: 0.8459466666666666
Testing accuracy: 0.83248
F1 score(train): 0.8477051643687555
F1 Score(test): 0.8331208160663054
```

	precision	recall	f1-score	support
0	0.84	0.82	0.83	6291
1	0.82	0.84	0.83	6209
accuracy			0.83	12500
macro avg	0.83	0.83	0.83	12500
weighted avg	0.83	0.83	0.83	12500

Confusion matrix which visualises and summarizes the performance of a classification problem.