Report on Sentiment Analysis of IMDB Movie Reviews

By

Syeda Maham Jafri (22796)

Abstract:
This report presents an extensive comparative analysis of sentiment analysis models applied to the IMDb dataset, integrating various preprocessing techniques and feature extraction methodologies. The study encompasses a wide array of machine learning algorithms including Naive Bayes, Decision Trees, Random Forest, Gradient Boosting, Stacking, Neural Networks, K-nearest neighbours, and Voting classifiers. Additionally, state-of-the-art pre-trained language models (PLMs) such as RoBERTa and DistilBERT and embedding techniques such as Glove (Global Vectors for Word Representation) and Word2Vec were employed to enhance feature representation. Furthermore, lexicon-based approaches such as AFINN, TextBlob, and VADER were incorporated for sentiment analysis. Diverse preprocessing techniques such as stop word removal, tokenization, POS tag removal, and different vectorization strategies including TF-IDF and count vectorizer were applied. Stemming techniques like Porter, Lancaster, and Snowball, as well as lemmatization via Spacy, were also utilized to preprocess the textual data. Through rigorous evaluation, the performance of each model and preprocessing combination is thoroughly assessed, providing detailed insights into the impact of preprocessing techniques and model choices on sentiment analysis accuracy.

Objective:
The objective of this assignment is to explore and compare various methods for text classification and sentiment analysis, ranging from traditional lexicon-based approaches to modern pre-trained Large Language Models (LLMs).

Dataset:
The IMDb dataset serves as a rich and diverse corpus for sentiment analysis tasks, comprising a vast collection of movie reviews contributed by users. With thousands of reviews spanning various genres, languages, and demographics, the dataset offers a comprehensive representation of audience opinions and sentiments towards movies. Each review within the IMDb dataset encapsulates valuable insights into viewers' perceptions, ranging from nuanced critiques to enthusiastic praises. Leveraging this extensive dataset, our analysis aims to uncover underlying sentiment trends and patterns, facilitating the evaluation and comparison of sentiment analysis models across diverse movie review scenarios. The training dataset (or portion of it) has been used to train the machine learning model, learning customized Word2Vec, or finetuning PLMs (RoBERTa, DistillBERT, GPT2). The test data has be used for evaluation purposes.

Preprocessing Techniques:

- **Stop Word Removal:**
  Stop words, such as "the," "is," and "and," were removed from the IMDb movie reviews. These words are common and do not contribute significant semantic meaning to the text, thus eliminating them can improve the efficiency and accuracy of sentiment analysis algorithms.

- **Tokenization:**
  The process of tokenization involved breaking down the movie reviews into individual words or tokens. This step facilitates further analysis by providing a structured representation of the text data, enabling algorithms to process and understand the content more effectively.

- **Part-of-Speech (POS) Tag Removal:**
  POS tags, indicating the grammatical category of each word (e.g., noun, verb, adjective), were removed from the movie reviews. This step simplified the text data while preserving essential semantic information, contributing to more accurate sentiment analysis results.

- **Vectorization**:
  Different vectorization techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) and count vectorizer were employed to convert the textual data into numerical representations. These representations enable machine learning algorithms to process and analyze the IMDb movie reviews effectively, capturing the underlying semantic information for sentiment analysis tasks.

- **Stemming and Lemmatization:**
  Stemming techniques including Porter, Lancaster, and Snowball, as well as lemmatization via Spacy, were applied to reduce words to their base or root form. This process standardizes word forms and reduces inflectional variants, ensuring consistency in the text data and improving sentiment analysis accuracy.

Model Evaluation:

Machine Learning Models

Stop words	Tokenize	Pos	Stemming/ Lemmatize	Vectorizer	Time Taken	Classification Report and Accuracy
yes	yes	yes	Porter Stemmer	Tf – Idf Vectorizer		Naive Bayes Accuracy: 0.8469 Classification Report for Naive Bayes: precision recall f1-score support negative 0.85 0.84 0.85 9935 positive 0.85 0.85 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 ['positive' 'positive' 'negative' ... 'negative' 'posit /n Random Forest Accuracy: 0.83375 Classification Report for Random Forest: precision recall f1-score support negative 0.83 0.84 0.83 9935 positive 0.84 0.82 0.83 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 k-NN Accuracy: 0.66965 Classification Report for k-NN: precision recall f1-score support negative 0.65 0.74 0.69 9935 positive 0.70 0.60 0.64 10065 accuracy 0.67 20000 macro avg 0.67 0.67 0.67 20000 weighted avg 0.67 0.67 0.67 20000 /n Decision Tree Accuracy: 0.69575 Classification Report for Decision Tree: precision recall f1-score support negative 0.69 0.70 0.70 9935 positive 0.70 0.69 0.70 10065 accuracy 0.70 20000 macro avg 0.70 0.70 0.70 20000 weighted avg 0.70 0.70 0.70 20000 /n Gradient Boosting Accuracy: 0.79635 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.83 0.75 0.78 9935 positive 0.77 0.84 0.81 10065 accuracy 0.80 20000 macro avg 0.80 0.80 0.80 20000 weighted avg 0.80 0.80 0.80 20000 Stacking Accuracy: 0.8613 Classification Report for Stacking: precision recall f1-score support negative 0.86 0.86 0.86 9935 positive 0.86 0.87 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000 Neural Network Accuracy: 0.8495 Classification Report for Neural Network: precision recall f1-score support negative 0.85 0.85 0.85 9935 positive 0.85 0.85 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 Bagging Accuracy: 0.7925 Classification Report for Bagging: precision recall f1-score support negative 0.79 0.79 0.79 9935 positive 0.79 0.79 0.79 10065 accuracy 0.79 20000 macro avg 0.79 0.79 0.79 20000 weighted avg 0.79 0.79 0.79 20000

yes	yes	yes	Porter Stemmer	Count Vectorizer		

```
Naive Bayes Accuracy: 0.83615
Classification Report for Naive Bayes:
              precision    recall  f1-score   support

    negative       0.83      0.84      0.84      9935
    positive       0.84      0.83      0.84     10065

    accuracy                           0.84     20000
   macro avg       0.84      0.84      0.84     20000
weighted avg       0.84      0.84      0.84     20000

['positive' 'positive' 'negative' ... 'negative' 'posit
/n
Random Forest Accuracy: 0.8348
Classification Report for Random Forest:
              precision    recall  f1-score   support

    negative       0.83      0.84      0.83      9935
    positive       0.84      0.83      0.83     10065

    accuracy                           0.83     20000
   macro avg       0.83      0.83      0.83     20000
weighted avg       0.83      0.83      0.83     20000

['positive' 'positive' 'negative' ... 'negative' 'posit
k-NN Accuracy: 0.58895
Classification Report for k-NN:
              precision    recall  f1-score   support

    negative       0.57      0.66      0.62      9935
    positive       0.61      0.52      0.56     10065

    accuracy                           0.59     20000
   macro avg       0.59      0.59      0.59     20000
weighted avg       0.59      0.59      0.59     20000

/n
Decision Tree Accuracy: 0.69715
Classification Report for Decision Tree:
              precision    recall  f1-score   support

    negative       0.70      0.69      0.69      9935
    positive       0.70      0.70      0.70     10065

    accuracy                           0.70     20000
   macro avg       0.70      0.70      0.70     20000
weighted avg       0.70      0.70      0.70     20000

/n
Gradient Boosting Accuracy: 0.79505
Classification Report for Gradient Boosting:
              precision    recall  f1-score   support

    negative       0.82      0.75      0.78      9935
    positive       0.77      0.84      0.81     10065

    accuracy                           0.80     20000
   macro avg       0.80      0.79      0.79     20000
weighted avg       0.80      0.80      0.79     20000

 Stacking Accuracy: 0.8534
 Classification Report for Stacking:
              precision    recall  f1-score   support

     negative       0.86      0.85      0.85      9935
     positive       0.85      0.86      0.86     10065

     accuracy                           0.85     20000
    macro avg       0.85      0.85      0.85     20000
 weighted avg       0.85      0.85      0.85     20000
Bagging Accuracy: 0.7902
Classification Report for Bagging:
              precision    recall  f1-score   support

    negative       0.79      0.78      0.79      9935
    positive       0.79      0.80      0.79     10065

    accuracy                           0.79     20000
   macro avg       0.79      0.79      0.79     20000
weighted avg       0.79      0.79      0.79     20000

/n
Voting Classifier Accuracy: 0.82985
Classification Report for Voting Classifier:
              precision    recall  f1-score   support

    negative       0.83      0.82      0.83      9935
    positive       0.83      0.84      0.83     10065

    accuracy                           0.83     20000
   macro avg       0.83      0.83      0.83     20000
weighted avg       0.83      0.83      0.83     20000

Neural Network Accuracy: 0.85705
Classification Report for Neural Network:
              precision    recall  f1-score   support

    negative       0.86      0.85      0.86      9935
    positive       0.85      0.86      0.86     10065

    accuracy                           0.86     20000
   macro avg       0.86      0.86      0.86     20000
weighted avg       0.86      0.86      0.86     20000
```

yes	yes	yes	Lancaster Stemmer	Tf – Idf Vectorizer		

```
Naive Bayes Accuracy: 0.845
Classification Report for Naive Bayes:
              precision    recall  f1-score   support

    negative       0.84      0.84      0.84      9935
    positive       0.85      0.85      0.85     10065

    accuracy                           0.84     20000
   macro avg       0.84      0.84      0.84     20000
weighted avg       0.84      0.84      0.84     20000

['positive' 'positive' 'negative' ... 'negative' 'posi
/n
Random Forest Accuracy: 0.83445
Classification Report for Random Forest:
              precision    recall  f1-score   support

    negative       0.83      0.84      0.83      9935
    positive       0.84      0.83      0.83     10065

    accuracy                           0.83     20000
   macro avg       0.83      0.83      0.83     20000
weighted avg       0.83      0.83      0.83     20000

['negative' 'positive' 'negative' ... 'negative' 'posi
k-NN Accuracy: 0.67225
Classification Report for k-NN:
              precision    recall  f1-score   support

    negative       0.64      0.77      0.70      9935
    positive       0.72      0.57      0.64     10065

    accuracy                           0.67     20000
   macro avg       0.68      0.67      0.67     20000
weighted avg       0.68      0.67      0.67     20000

/n
Decision Tree Accuracy: 0.6931
Classification Report for Decision Tree:
              precision    recall  f1-score   support

    negative       0.69      0.70      0.69      9935
    positive       0.70      0.69      0.69     10065

    accuracy                           0.69     20000
   macro avg       0.69      0.69      0.69     20000
weighted avg       0.69      0.69      0.69     20000

/n
Gradient Boosting Accuracy: 0.79635
Classification Report for Gradient Boosting:
              precision    recall  f1-score   support

    negative       0.82      0.75      0.79      9935
    positive       0.77      0.84      0.81     10065

    accuracy                           0.80     20000
   macro avg       0.80      0.80      0.80     20000
weighted avg       0.80      0.80      0.80     20000

Stacking Accuracy: 0.8602
Classification Report for Stacking:
              precision    recall  f1-score   support

    negative       0.87      0.85      0.86      9935
    positive       0.86      0.87      0.86     10065

    accuracy                           0.86     20000
   macro avg       0.86      0.86      0.86     20000
weighted avg       0.86      0.86      0.86     20000

Bagging Accuracy: 0.79065
Classification Report for Bagging:
              precision    recall  f1-score   support

    negative       0.79      0.78      0.79      9935
    positive       0.79      0.80      0.79     10065

    accuracy                           0.79     20000
   macro avg       0.79      0.79      0.79     20000
weighted avg       0.79      0.79      0.79     20000

/n
Voting Classifier Accuracy: 0.8408
Classification Report for Voting Classifier:
              precision    recall  f1-score   support

    negative       0.84      0.84      0.84      9935
    positive       0.84      0.84      0.84     10065

    accuracy                           0.84     20000
   macro avg       0.84      0.84      0.84     20000
weighted avg       0.84      0.84      0.84     20000

Neural Network Accuracy: 0.84925
Classification Report for Neural Network:
              precision    recall  f1-score   support

    negative       0.85      0.84      0.85      9935
    positive       0.85      0.85      0.85     10065

    accuracy                           0.85     20000
   macro avg       0.85      0.85      0.85     20000
weighted avg       0.85      0.85      0.85     20000
```

yes	yes	yes	Lancaster Stemmer	Count Vectorizer		Naive Bayes Accuracy: 0.8342 Classification Report for Naive Bayes: precision recall f1-score support negative 0.83 0.84 0.83 9935 positive 0.84 0.83 0.83 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 ['positive' 'positive' 'negative' ... 'negative' 'posi /n Random Forest Accuracy: 0.8292 Classification Report for Random Forest: precision recall f1-score support negative 0.83 0.83 0.83 9935 positive 0.83 0.83 0.83 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 ['negative' 'positive' 'negative' ... 'negative' 'posi Neural Network Accuracy: 0.85775 Classification Report for Neural Network: precision recall f1-score support negative 0.86 0.85 0.86 9935 positive 0.85 0.87 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000 Decision Tree Accuracy: 0.6954 Classification Report for Decision Tree: precision recall f1-score support negative 0.69 0.70 0.69 9935 positive 0.70 0.69 0.70 10065 accuracy 0.70 20000 macro avg 0.70 0.70 0.70 20000 weighted avg 0.70 0.70 0.70 20000 /n Gradient Boosting Accuracy: 0.79555 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.82 0.75 0.78 9935 positive 0.77 0.84 0.81 10065 accuracy 0.80 20000 macro avg 0.80 0.80 0.79 20000 weighted avg 0.80 0.80 0.80 20000 /n Stacking Accuracy: 0.8516 Classification Report for Stacking: precision recall f1-score support negative 0.86 0.84 0.85 9935 positive 0.85 0.86 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 Bagging Accuracy: 0.79035 Classification Report for Bagging: precision recall f1-score support negative 0.79 0.78 0.79 9935 positive 0.79 0.80 0.79 10065 accuracy 0.79 20000 macro avg 0.79 0.79 0.79 20000 weighted avg 0.79 0.79 0.79 20000 /n Voting Classifier Accuracy: 0.82985 Classification Report for Voting Classifier: precision recall f1-score support negative 0.83 0.82 0.83 9935 positive 0.83 0.84 0.83 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000

yes	yes		Lancaster Stemmer	Count Vectorizer		Naive Bayes Accuracy: 0.8427 Classification Report for Naive Bayes: precision recall f1-score support negative 0.83 0.85 0.84 9935 positive 0.85 0.83 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 ['positive' 'positive' 'negative' ... 'negative' 'posi /n Random Forest Accuracy: 0.8427 Classification Report for Random Forest: precision recall f1-score support negative 0.84 0.85 0.84 9935 positive 0.85 0.84 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 ['negative' 'positive' 'negative' ... 'negative' 'posi Neural Network Accuracy: 0.86675 Classification Report for Neural Network: precision recall f1-score support negative 0.87 0.86 0.86 9935 positive 0.86 0.88 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 Decision Tree Accuracy: 0.723 Classification Report for Decision Tree: precision recall f1-score support negative 0.72 0.72 0.72 9935 positive 0.73 0.72 0.72 10065 accuracy 0.72 20000 macro avg 0.72 0.72 0.72 20000 weighted avg 0.72 0.72 0.72 20000 /n Gradient Boosting Accuracy: 0.8084 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.84 0.76 0.80 9935 positive 0.79 0.85 0.82 10065 accuracy 0.81 20000 macro avg 0.81 0.81 0.81 20000 weighted avg 0.81 0.81 0.81 20000 k-NN Accuracy: 0.62795 Classification Report for k-NN: precision recall f1-score support negative 0.61 0.71 0.65 9935 positive 0.66 0.55 0.60 10065 accuracy 0.63 20000 macro avg 0.63 0.63 0.63 20000 weighted avg 0.63 0.63 0.63 20000 /n Stacking Accuracy: 0.8634 Classification Report for Stacking: precision recall f1-score support negative 0.86 0.86 0.86 9935 positive 0.87 0.86 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000 Bagging Accuracy: 0.80275 Classification Report for Bagging: precision recall f1-score support negative 0.81 0.79 0.80 9935 positive 0.80 0.82 0.81 10065 accuracy 0.80 20000 macro avg 0.80 0.80 0.80 20000 weighted avg 0.80 0.80 0.80 20000 /n Voting Classifier Accuracy: 0.84365 Classification Report for Voting Classifier: precision recall f1-score support negative 0.84 0.84 0.84 9935 positive 0.84 0.85 0.85 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000

yes	yes		Lancaster Stemmer	Count Vectorizer		**Modified ML Parameters** MultinomialNB(alpha=0.1) Naive Bayes Accuracy: 0.84295 Classification Report for Naive Bayes: precision recall f1-score support negative 0.84 0.85 0.84 9935 positive 0.85 0.83 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 RandomForestClassifier(n_estimators=200, max_depth=10, random_state=42) Random Forest Accuracy: 0.8277 Classification Report for Random Forest: precision recall f1-score support negative 0.86 0.78 0.82 9935 positive 0.80 0.87 0.84 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 MLPClassifier(hidden_layer_sizes=(128,), max_iter=200, alpha=0.0001, random_state=42) Neural Network MLP CLASSIFIER Accuracy: 0.8717 Classification Report for Neural Network: precision recall f1-score support negative 0.88 0.86 0.87 9935 positive 0.87 0.88 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 DecisionTreeClassifier(max_depth=10, min_samples_split=10, random_state=42) Decision Tree Accuracy: 0.736 Classification Report for Decision Tree: precision recall f1-score support negative 0.78 0.65 0.71 9935 positive 0.70 0.82 0.76 10065 accuracy 0.74 20000 macro avg 0.74 0.74 0.73 20000 weighted avg 0.74 0.74 0.73 20000 GradientBoostingClassifier(n_estimators=200, max_depth=5, learning_rate=0.1, random_state=42) Gradient Boosting Accuracy: 0.8446 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.86 0.83 0.84 9935 positive 0.83 0.86 0.85 10065 accuracy 0.84 20000 macro avg 0.85 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 KNeighborsClassifier(n_neighbors=10) k-NN Accuracy: 0.6297 Classification Report for k-NN: precision recall f1-score support negative 0.59 0.80 0.68 9935 positive 0.70 0.46 0.55 10065 accuracy 0.63 20000 macro avg 0.65 0.63 0.62 20000 weighted avg 0.65 0.63 0.62 20000 StackingClassifier(estimators=[('nb', nb_model), ('rf', rf_model), ('knn', knn_model), ('dt', dt_model),('gb', gb_model)], final_estimator=GradientBoostingClassifier(n_estimators=100, random_state=42)) Stacking Accuracy: 0.869 Classification Report for Stacking: precision recall f1-score support negative 0.87 0.87 0.87 9935 positive 0.87 0.87 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000

yes	yes	yes	Lancaster Stemmer	Tf-Idf Vectorizer + robust Scaler		Naive Bayes Accuracy: 0.84545 Classification Report for Naive Bayes: precision recall f1-score support negative 0.85 0.84 0.84 9935 positive 0.84 0.85 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 Random Forest Accuracy: 0.8216 Classification Report for Random Forest: precision recall f1-score support negative 0.84 0.79 0.81 9935 positive 0.80 0.86 0.83 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000 Neural Network MLP CLASSIFIER Accuracy: 0.85495 Classification Report for Neural Network: precision recall f1-score support negative 0.86 0.85 0.85 9935 positive 0.85 0.86 0.86 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 Decision Tree Accuracy: 0.72305 Classification Report for Decision Tree: precision recall f1-score support negative 0.78 0.62 0.69 9935 positive 0.69 0.83 0.75 10065 accuracy 0.72 20000 macro avg 0.73 0.72 0.72 20000 weighted avg 0.73 0.72 0.72 20000 Gradient Boosting Accuracy: 0.8319 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.85 0.81 0.83 9935 positive 0.82 0.86 0.84 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 k-NN Accuracy: 0.6698 Classification Report for k-NN: precision recall f1-score support negative 0.62 0.86 0.72 9935 positive 0.78 0.48 0.59 10065 accuracy 0.67 20000 macro avg 0.70 0.67 0.66 20000 weighted avg 0.70 0.67 0.66 20000
yes	yes	yes	Lancaster Stemmer	Tf-Idf Vectorizer + MinMax Scaler		Naive Bayes Accuracy: 0.84465 Classification Report for Naive Bayes: precision recall f1-score support negative 0.85 0.84 0.84 9935 positive 0.84 0.85 0.85 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 Random Forest Accuracy: 0.8219 Classification Report for Random Forest: precision recall f1-score support negative 0.84 0.79 0.81 9935 positive 0.80 0.86 0.83 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000

						Neural Network MLP CLASSIFIER Accuracy: 0.85705 Classification Report for Neural Network: precision recall f1-score support negative 0.86 0.85 0.86 9935 positive 0.86 0.86 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000 Decision Tree Accuracy: 0.72305 Classification Report for Decision Tree: precision recall f1-score support negative 0.78 0.62 0.69 9935 positive 0.69 0.83 0.75 10065 accuracy 0.72 20000 macro avg 0.73 0.72 0.72 20000 weighted avg 0.73 0.72 0.72 20000 Gradient Boosting Accuracy: 0.8305 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.85 0.80 0.83 9935 positive 0.82 0.86 0.84 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 k-NN Accuracy: 0.501 Classification Report for k-NN: precision recall f1-score support negative 0.50 0.99 0.66 9935 positive 0.69 0.02 0.03 10065 accuracy 0.50 20000 macro avg 0.60 0.50 0.35 20000 weighted avg 0.60 0.50 0.34 20000
yes	yes	yes	Lancaster Stemmer	Tf-Idf Vectorizer + Standard Scaler		Naive Bayes Accuracy: 0.84545 Classification Report for Naive Bayes: precision recall f1-score support negative 0.85 0.84 0.84 9935 positive 0.84 0.85 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 Random Forest Accuracy: 0.8216 Classification Report for Random Forest: precision recall f1-score support negative 0.84 0.79 0.81 9935 positive 0.80 0.86 0.83 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000 Neural Network MLP CLASSIFIER Accuracy: 0.85495 Classification Report for Neural Network: precision recall f1-score support negative 0.86 0.85 0.85 9935 positive 0.85 0.86 0.86 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 k-NN Accuracy: 0.6698 Classification Report for k-NN: precision recall f1-score support negative 0.62 0.86 0.72 9935 positive 0.78 0.48 0.59 10065 accuracy 0.67 20000 macro avg 0.70 0.67 0.66 20000 weighted avg 0.70 0.67 0.66 20000 Stacking Accuracy: 0.8646 Classification Report for Stacking: precision recall f1-score support negative 0.87 0.85 0.86 9935 positive 0.86 0.88 0.87 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000

yes	yes	yes	Lancaster Stemmer	Tf-Idf Vectorizer + Standard Scaler		Decision Tree Accuracy: 0.72305 Classification Report for Decision Tree: precision recall f1-score support negative 0.78 0.62 0.69 9935 positive 0.69 0.83 0.75 10065 accuracy 0.72 20000 macro avg 0.73 0.72 0.72 20000 weighted avg 0.73 0.72 0.72 20000 Gradient Boosting Accuracy: 0.8319 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.85 0.81 0.83 9935 positive 0.82 0.86 0.84 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000
yes	yes	yes	Snowball Stemmer	TF-IDF Vectorizer		Naive Bayes Accuracy: 0.84435 Classification Report for Naive Bayes: precision recall f1-score support negative 0.85 0.84 0.84 9935 positive 0.84 0.85 0.85 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 ['positive' 'positive' 'negative' ... 'negative' 'posi /n Random Forest Accuracy: 0.83315 Classification Report for Random Forest: precision recall f1-score support negative 0.83 0.84 0.83 9935 positive 0.84 0.83 0.83 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 ['positive' 'positive' 'negative' ... 'negative' 'posi Neural Network Accuracy: 0.84755 Classification Report for Neural Network: precision recall f1-score support negative 0.85 0.85 0.85 9935 positive 0.85 0.85 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 Decision Tree Accuracy: 0.69425 Classification Report for Decision Tree: precision recall f1-score support negative 0.69 0.70 0.70 9935 positive 0.70 0.68 0.69 10065 accuracy 0.69 20000 macro avg 0.69 0.69 0.69 20000 weighted avg 0.69 0.69 0.69 20000 Gradient Boosting Accuracy: 0.79755 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.83 0.75 0.79 9935 positive 0.77 0.84 0.81 10065 accuracy 0.80 20000 macro avg 0.80 0.80 0.80 20000 weighted avg 0.80 0.80 0.80 20000 k-NN Accuracy: 0.6682 Classification Report for k-NN: precision recall f1-score support negative 0.65 0.73 0.69 9935 positive 0.70 0.60 0.65 10065 accuracy 0.67 20000 macro avg 0.67 0.67 0.67 20000 weighted avg 0.67 0.67 0.67 20000 /n Voting Classifier Accuracy: 0.841 Classification Report for Voting Classifier: precision recall f1-score support negative 0.84 0.83 0.84 9935 positive 0.84 0.85 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000

yes	yes	yes	Snowball Stemmer	TF-IDF Vectorizer		Stacking Accuracy: 0.8614 Classification Report for Stacking: precision recall f1-score support negative 0.86 0.86 0.86 9935 positive 0.86 0.86 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000 /n Bagging Accuracy: 0.7939 Classification Report for Bagging: precision recall f1-score support negative 0.79 0.79 0.79 9935 positive 0.79 0.80 0.80 10065 accuracy 0.79 20000 macro avg 0.79 0.79 0.79 20000 weighted avg 0.79 0.79 0.79 20000
yes	yes	yes	Snowball Stemmer	TF-IDF Vectorizer + MinMax Scaler		Naive Bayes Accuracy: 0.8446 Classification Report for Naive Bayes: precision recall f1-score support negative 0.85 0.84 0.84 9935 positive 0.84 0.85 0.85 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 Random Forest Accuracy: 0.8209 Classification Report for Random Forest: precision recall f1-score support negative 0.85 0.78 0.81 9935 positive 0.80 0.86 0.83 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000 Neural Network MLP CLASSIFIER Accuracy: 0.85665 Classification Report for Neural Network: precision recall f1-score support negative 0.86 0.86 0.86 9935 positive 0.86 0.86 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000 Decision Tree Accuracy: 0.7263 Classification Report for Decision Tree: precision recall f1-score support negative 0.79 0.61 0.69 9935 positive 0.69 0.84 0.76 10065 accuracy 0.73 20000 macro avg 0.74 0.73 0.72 20000 weighted avg 0.74 0.73 0.72 20000 Gradient Boosting Accuracy: 0.83305 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.85 0.81 0.83 9935 positive 0.82 0.86 0.84 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 k-NN Accuracy: 0.50195 Classification Report for k-NN: precision recall f1-score support negative 0.50 1.00 0.67 9935 positive 0.77 0.01 0.03 10065 accuracy 0.50 20000 macro avg 0.63 0.51 0.35 20000 weighted avg 0.63 0.50 0.34 20000

yes	yes		Snowball Stemmer	Tf – Idf Vectorizer		Naive Bayes Accuracy: 0.85145 Classification Report for Naive Bayes: precision recall f1-score support negative 0.85 0.85 0.85 9935 positive 0.85 0.85 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 ['positive' 'positive' 'negative' ... 'negative' 'posit /n Random Forest Accuracy: 0.8484 Classification Report for Random Forest: precision recall f1-score support negative 0.84 0.86 0.85 9935 positive 0.86 0.84 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 k-NN Accuracy: 0.74325 Classification Report for k-NN: precision recall f1-score support negative 0.76 0.71 0.73 9935 positive 0.73 0.77 0.75 10065 accuracy 0.74 20000 macro avg 0.74 0.74 0.74 20000 weighted avg 0.74 0.74 0.74 20000 /n Decision Tree Accuracy: 0.72035 Classification Report for Decision Tree: precision recall f1-score support negative 0.72 0.72 0.72 9935 positive 0.72 0.72 0.72 10065 accuracy 0.72 20000 macro avg 0.72 0.72 0.72 20000 weighted avg 0.72 0.72 0.72 20000 /n Gradient Boosting Accuracy: 0.81575 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.84 0.77 0.81 9935 positive 0.79 0.86 0.82 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000 Neural Network Accuracy: 0.8579 Classification Report for Neural Network: precision recall f1-score support negative 0.86 0.85 0.86 9935 positive 0.86 0.86 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000 Stacking Accuracy: 0.87435 Classification Report for Stacking: precision recall f1-score support negative 0.87 0.87 0.87 9935 positive 0.88 0.88 0.88 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 Bagging Accuracy: 0.8092 Classification Report for Bagging: precision recall f1-score support negative 0.81 0.81 0.81 9935 positive 0.81 0.81 0.81 10065 accuracy 0.81 20000 macro avg 0.81 0.81 0.81 20000 weighted avg 0.81 0.81 0.81 20000 /n Voting Classifier Accuracy: 0.8566 Classification Report for Voting Classifier: precision recall f1-score support negative 0.87 0.84 0.85 9935 positive 0.85 0.87 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000

yes	yes	yes	Snowball Stemmer	Count Vectorizer		Neural Network Accuracy: 0.8581 Classification Report for Neural Network: precision recall f1-score support negative 0.86 0.85 0.86 9935 positive 0.86 0.86 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000 Decision Tree Accuracy: 0.7009 Classification Report for Decision Tree: precision recall f1-score support negative 0.70 0.70 0.70 9935 positive 0.70 0.71 0.70 10065 accuracy 0.70 20000 macro avg 0.70 0.70 0.70 20000 weighted avg 0.70 0.70 0.70 20000 Gradient Boosting Accuracy: 0.79745 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.83 0.75 0.79 9935 positive 0.77 0.85 0.81 10065 accuracy 0.80 20000 macro avg 0.80 0.80 0.80 20000 weighted avg 0.80 0.80 0.80 20000 Naive Bayes Accuracy: 0.83735 Classification Report for Naive Bayes: precision recall f1-score support negative 0.83 0.84 0.84 9935 positive 0.84 0.83 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 Random Forest Accuracy: 0.83265 Classification Report for Random Forest: precision recall f1-score support negative 0.83 0.84 0.83 9935 positive 0.84 0.82 0.83 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 k-NN Accuracy: 0.59095 Classification Report for k-NN: precision recall f1-score support negative 0.58 0.67 0.62 9935 positive 0.61 0.51 0.56 10065 accuracy 0.59 20000 macro avg 0.59 0.59 0.59 20000 weighted avg 0.59 0.59 0.59 20000 Stacking Accuracy: 0.85405 Classification Report for Stacking: precision recall f1-score support negative 0.86 0.85 0.85 9935 positive 0.85 0.86 0.86 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 Bagging Accuracy: 0.7924 Classification Report for Bagging: precision recall f1-score support negative 0.80 0.78 0.79 9935 positive 0.79 0.81 0.80 10065 accuracy 0.79 20000 macro avg 0.79 0.79 0.79 20000 weighted avg 0.79 0.79 0.79 20000 Voting Classifier Accuracy: 0.83145 Classification Report for Voting Classifier: precision recall f1-score support negative 0.84 0.82 0.83 9935 positive 0.83 0.84 0.83 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000

yes	yes		Snowball Stemmer	Count Vectorizer		Naive Bayes Accuracy: 0.84445 Classification Report for Naive Bayes: precision recall f1-score support negative 0.84 0.85 0.84 9935 positive 0.85 0.84 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 Random Forest Accuracy: 0.8472 Classification Report for Random Forest: precision recall f1-score support negative 0.84 0.85 0.85 9935 positive 0.85 0.84 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 Neural Network Accuracy: 0.8672 Classification Report for Neural Network: precision recall f1-score support negative 0.87 0.86 0.87 9935 positive 0.86 0.87 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 Decision Tree Accuracy: 0.7212 Classification Report for Decision Tree: precision recall f1-score support negative 0.72 0.72 0.72 9935 positive 0.72 0.72 0.72 10065 accuracy 0.72 20000 macro avg 0.72 0.72 0.72 20000 weighted avg 0.72 0.72 0.72 20000 Gradient Boosting Accuracy: 0.81135 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.84 0.77 0.80 9935 positive 0.79 0.85 0.82 10065 accuracy 0.81 20000 macro avg 0.81 0.81 0.81 20000 weighted avg 0.81 0.81 0.81 20000 k-NN Accuracy: 0.6262 Classification Report for k-NN: precision recall f1-score support negative 0.61 0.68 0.64 9935 positive 0.64 0.57 0.61 10065 accuracy 0.63 20000 macro avg 0.63 0.63 0.63 20000 weighted avg 0.63 0.63 0.63 20000 Stacking Accuracy: 0.86565 Classification Report for Stacking: precision recall f1-score support negative 0.86 0.87 0.87 9935 positive 0.87 0.86 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 Bagging Accuracy: 0.80165 Classification Report for Bagging: precision recall f1-score support negative 0.81 0.79 0.80 9935 positive 0.80 0.81 0.81 10065 accuracy 0.80 20000 macro avg 0.80 0.80 0.80 20000 weighted avg 0.80 0.80 0.80 20000 Voting Classifier Accuracy: 0.84455 Classification Report for Voting Classifier: precision recall f1-score support negative 0.85 0.83 0.84 9935 positive 0.84 0.86 0.85 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000

yes	yes		Lancaster Stemmer	Tf-Idf Vectorizer		Naive Bayes Accuracy: 0.8503 Classification Report for Naive Bayes: precision recall f1-score support negative 0.85 0.85 0.85 9935 positive 0.85 0.85 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 Random Forest Accuracy: 0.84525 Classification Report for Random Forest: precision recall f1-score support negative 0.84 0.85 0.85 9935 positive 0.85 0.84 0.84 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 Neural Network Accuracy: 0.8533 Classification Report for Neural Network: precision recall f1-score support negative 0.85 0.85 0.85 9935 positive 0.85 0.86 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 Decision Tree Accuracy: 0.7125 Classification Report for Decision Tree: precision recall f1-score support negative 0.71 0.72 0.71 9935 positive 0.72 0.71 0.71 10065 accuracy 0.71 20000 macro avg 0.71 0.71 0.71 20000 weighted avg 0.71 0.71 0.71 20000 Gradient Boosting Accuracy: 0.81185 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.84 0.77 0.80 9935 positive 0.79 0.85 0.82 10065 accuracy 0.81 20000 macro avg 0.81 0.81 0.81 20000 weighted avg 0.81 0.81 0.81 20000 k-NN Accuracy: 0.74355 Classification Report for k-NN: precision recall f1-score support negative 0.75 0.72 0.74 9935 positive 0.74 0.77 0.75 10065 accuracy 0.74 20000 macro avg 0.74 0.74 0.74 20000 weighted avg 0.74 0.74 0.74 20000 Stacking Accuracy: 0.87125 Classification Report for Stacking: precision recall f1-score support negative 0.87 0.87 0.87 9935 positive 0.87 0.87 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 Bagging Accuracy: 0.803 Classification Report for Bagging: precision recall f1-score support negative 0.81 0.80 0.80 9935 positive 0.80 0.81 0.81 10065 accuracy 0.80 20000 macro avg 0.80 0.80 0.80 20000 weighted avg 0.80 0.80 0.80 20000 Voting Classifier Accuracy: 0.85565 Classification Report for Voting Classifier: precision recall f1-score support negative 0.86 0.84 0.85 9935 positive 0.85 0.87 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000
yes	yes	yes	Spacy	Tf – Idf Vectorizer		Naive Bayes Accuracy: 0.8513 Classification Report for Naive Bayes: precision recall f1-score support negative 0.85 0.85 0.85 9935 positive 0.85 0.85 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000

						Random Forest Accuracy: 0.8443 Classification Report for Random Forest: precision recall f1-score support negative 0.84 0.85 0.85 9935 positive 0.85 0.83 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 Neural Network MLP CLASSIFIER Accuracy: 0.85735 Classification Report for Neural Network: precision recall f1-score support negative 0.85 0.86 0.86 9935 positive 0.86 0.85 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000 Decision Tree Accuracy: 0.7083 Classification Report for Decision Tree: precision recall f1-score support negative 0.70 0.71 0.71 9935 positive 0.71 0.71 0.71 10065 accuracy 0.71 20000 macro avg 0.71 0.71 0.71 20000 weighted avg 0.71 0.71 0.71 20000 Gradient Boosting Accuracy: 0.80335 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.83 0.76 0.79 9935 positive 0.78 0.85 0.81 10065 accuracy 0.80 20000 macro avg 0.81 0.80 0.80 20000 weighted avg 0.81 0.80 0.80 20000 k-NN Accuracy: 0.60415 Classification Report for k-NN: precision recall f1-score support negative 0.62 0.54 0.58 9935 positive 0.60 0.67 0.63 10065 accuracy 0.60 20000 macro avg 0.61 0.60 0.60 20000 weighted avg 0.61 0.60 0.60 20000 Stacking Accuracy: 0.865 Classification Report for Stacking: precision recall f1-score support negative 0.86 0.87 0.86 9935 positive 0.87 0.86 0.87 10065 accuracy 0.86 20000 macro avg 0.86 0.87 0.86 20000 weighted avg 0.87 0.86 0.87 20000 Bagging Accuracy: 0.8038 Classification Report for Bagging: precision recall f1-score support negative 0.80 0.81 0.80 9935 positive 0.81 0.80 0.80 10065 accuracy 0.80 20000 macro avg 0.80 0.80 0.80 20000 weighted avg 0.80 0.80 0.80 20000 Voting Classifier Accuracy: 0.84105 Classification Report for Voting Classifier: precision recall f1-score support negative 0.85 0.82 0.84 9935 positive 0.83 0.86 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000
yes	yes		Spacy	Tf – Idf Vectorizer		Naive Bayes Accuracy: 0.85375 Classification Report for Naive Bayes: precision recall f1-score support negative 0.85 0.85 0.85 9935 positive 0.85 0.85 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 Random Forest Accuracy: 0.8441 Classification Report for Random Forest: precision recall f1-score support negative 0.83 0.86 0.85 9935 positive 0.85 0.83 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000

						Neural Network MLP CLASSIFIER Accuracy: 0.858 Classification Report for Neural Network: precision recall f1-score support negative 0.86 0.86 0.86 9935 positive 0.86 0.86 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000 Decision Tree Accuracy: 0.717 Classification Report for Decision Tree: precision recall f1-score support negative 0.71 0.72 0.72 9935 positive 0.72 0.71 0.72 10065 accuracy 0.72 20000 macro avg 0.72 0.72 0.72 20000 weighted avg 0.72 0.72 0.72 20000 Gradient Boosting Accuracy: 0.81005 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.84 0.76 0.80 9935 positive 0.79 0.86 0.82 10065 accuracy 0.81 20000 macro avg 0.81 0.81 0.81 20000 weighted avg 0.81 0.81 0.81 20000 k-NN Accuracy: 0.73815 Classification Report for k-NN: precision recall f1-score support negative 0.75 0.70 0.73 9935 positive 0.72 0.77 0.75 10065 accuracy 0.74 20000 macro avg 0.74 0.74 0.74 20000 weighted avg 0.74 0.74 0.74 20000 Stacking Accuracy: 0.87125 Classification Report for Stacking: precision recall f1-score support negative 0.87 0.87 0.87 9935 positive 0.87 0.87 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 Bagging Accuracy: 0.80865 Classification Report for Bagging: precision recall f1-score support negative 0.81 0.81 0.81 9935 positive 0.81 0.81 0.81 10065 accuracy 0.81 20000 macro avg 0.81 0.81 0.81 20000 weighted avg 0.81 0.81 0.81 20000 Voting Classifier Accuracy: 0.8546 Classification Report for Voting Classifier: precision recall f1-score support negative 0.86 0.84 0.85 9935 positive 0.85 0.87 0.86 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000
yes	yes		Spacy	Tf – Idf Vectorizer		Modified ML Parameter MultinomialNB(alpha=0.1) Naive Bayes Accuracy: 0.8535 Classification Report for Naive Bayes: precision recall f1-score support negative 0.85 0.85 0.85 9935 positive 0.85 0.86 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 RandomForestClassifier(n_estimators=200, max_depth=10, random_state=42) Random Forest Accuracy: 0.82955 Classification Report for Random Forest: precision recall f1-score support negative 0.85 0.80 0.82 9935 positive 0.81 0.86 0.84 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 nn_model = MLPClassifier(hidden_layer_sizes=(128,), max_iter=200, alpha=0.0001, random_state=42)

						Neural Network MLP CLASSIFIER Accuracy: 0.8603 Classification Report for Neural Network: precision recall f1-score support negative 0.86 0.86 0.86 9935 positive 0.86 0.86 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000 DecisionTreeClassifier(max_depth=10, min_samples_split=10, random_state=42) Decision Tree Accuracy: 0.7237 Classification Report for Decision Tree: precision recall f1-score support negative 0.72 0.73 0.72 9935 positive 0.73 0.72 0.72 10065 accuracy 0.72 20000 macro avg 0.72 0.72 0.72 20000 weighted avg 0.72 0.72 0.72 20000 GradientBoostingClassifier(n_estimators=200, max_depth=5, learning_rate=0.1, random_state=42) Gradient Boosting Accuracy: 0.84485 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.86 0.82 0.84 9935 positive 0.83 0.87 0.85 10065 accuracy 0.84 20000 macro avg 0.85 0.84 0.84 20000 weighted avg 0.85 0.84 0.84 20000 KNeighborsClassifier(n_neighbors=10) k-NN Accuracy: 0.7522 Classification Report for k-NN: precision recall f1-score support negative 0.73 0.80 0.76 9935 positive 0.78 0.71 0.74 10065 accuracy 0.75 20000 macro avg 0.75 0.75 0.75 20000 weighted avg 0.75 0.75 0.75 20000 StackingClassifier(estimators=[('nb', nb_model), ('rf', rf_model), ('knn', knn_model), ('dt', dt_model),('gb', gb_model)], final_estimator=GradientBoostingClassifier(n_estimators=100, random_state=42)) Stacking Accuracy: 0.8773 Classification Report for Stacking: precision recall f1-score support negative 0.88 0.88 0.88 9935 positive 0.88 0.88 0.88 10065 accuracy 0.88 20000 macro avg 0.88 0.88 0.88 20000 weighted avg 0.88 0.88 0.88 20000
yes	yes		Spacy	Count Vectorizer		Naive Bayes Accuracy: 0.84485 Classification Report for Naive Bayes: precision recall f1-score support negative 0.84 0.85 0.85 9935 positive 0.85 0.84 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 Random Forest Accuracy: 0.8448 Classification Report for Random Forest: precision recall f1-score support negative 0.84 0.85 0.85 9935 positive 0.85 0.84 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000

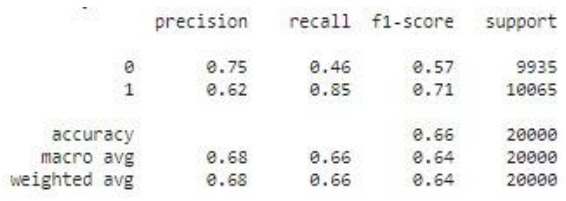
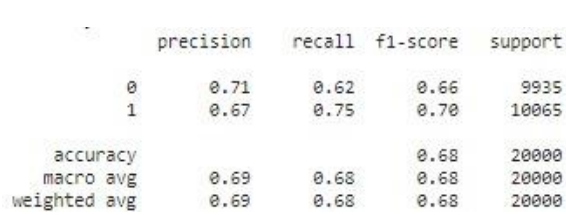
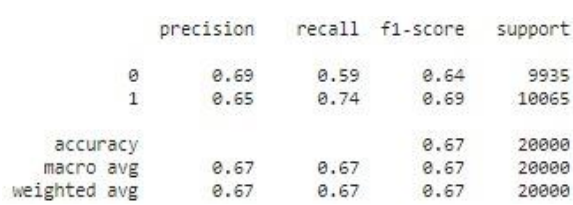
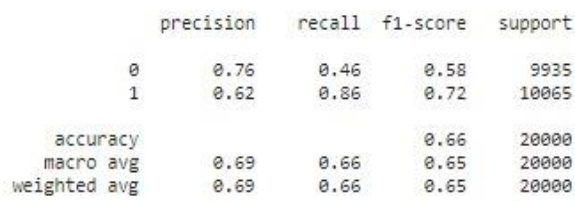
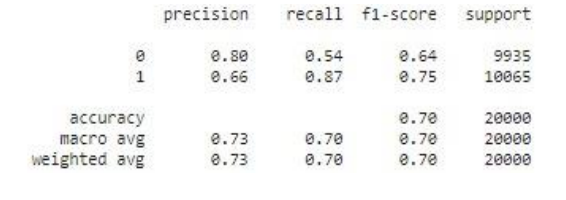
						Neural Network MLP CLASSIFIER Accuracy: 0.8652 Classification Report for Neural Network: precision recall f1-score support negative 0.87 0.86 0.86 9935 positive 0.86 0.87 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 Decision Tree Accuracy: 0.7161 Classification Report for Decision Tree: precision recall f1-score support negative 0.71 0.73 0.72 9935 positive 0.72 0.70 0.71 10065 accuracy 0.72 20000 macro avg 0.72 0.72 0.72 20000 weighted avg 0.72 0.72 0.72 20000 Gradient Boosting Accuracy: 0.8064 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.84 0.76 0.80 9935 positive 0.78 0.85 0.82 10065 accuracy 0.81 20000 macro avg 0.81 0.81 0.81 20000 weighted avg 0.81 0.81 0.81 20000 k-NN Accuracy: 0.64195 Classification Report for k-NN: precision recall f1-score support negative 0.62 0.70 0.66 9935 positive 0.66 0.58 0.62 10065 accuracy 0.64 20000 macro avg 0.64 0.64 0.64 20000 weighted avg 0.64 0.64 0.64 20000 Stacking Accuracy: 0.86475 Classification Report for Stacking: precision recall f1-score support negative 0.86 0.87 0.86 9935 positive 0.87 0.86 0.87 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000 Bagging Accuracy: 0.8032 Classification Report for Bagging: precision recall f1-score support negative 0.81 0.79 0.80 9935 positive 0.80 0.81 0.81 10065 accuracy 0.80 20000 macro avg 0.80 0.80 0.80 20000 weighted avg 0.80 0.80 0.80 20000 Voting Classifier Accuracy: 0.8439 Classification Report for Voting Classifier: precision recall f1-score support negative 0.85 0.84 0.84 9935 positive 0.84 0.85 0.85 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000
yes	yes		Spacy	Count Vectorizer		**Modified ML Parameters** Naive Bayes Accuracy: 0.8446 Classification Report for Naive Bayes: precision recall f1-score support negative 0.84 0.85 0.84 9935 positive 0.85 0.84 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 Random Forest Accuracy: 0.82905 Classification Report for Random Forest: precision recall f1-score support negative 0.85 0.80 0.82 9935 positive 0.81 0.86 0.83 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000

						Neural Network MLP CLASSIFIER Accuracy: 0.8728 Classification Report for Neural Network: precision recall f1-score support negative 0.88 0.87 0.87 9935 positive 0.87 0.88 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 Decision Tree Accuracy: 0.7378 Classification Report for Decision Tree: precision recall f1-score support negative 0.79 0.64 0.71 9935 positive 0.70 0.83 0.76 10065 accuracy 0.74 20000 macro avg 0.75 0.74 0.74 20000 weighted avg 0.75 0.74 0.74 20000 Gradient Boosting Accuracy: 0.84685 Classification Report for Gradient Boosting: precision recall f1-score support negative 0.86 0.82 0.84 9935 positive 0.83 0.87 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 k-NN Accuracy: 0.64515 Classification Report for k-NN: precision recall f1-score support negative 0.61 0.79 0.69 9935 positive 0.71 0.50 0.59 10065 accuracy 0.65 20000 macro avg 0.66 0.65 0.64 20000 weighted avg 0.66 0.65 0.64 20000 Stacking Accuracy: 0.86975 Classification Report for Stacking: precision recall f1-score support negative 0.87 0.87 0.87 9935 positive 0.87 0.87 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000

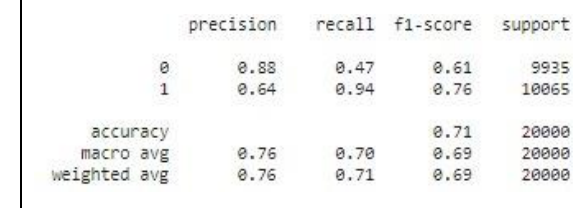
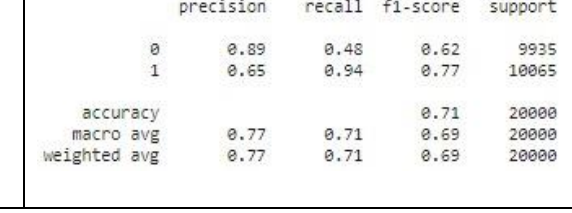
Lexicon Based Approaches

1- AFINN

Stop word	Tokenize	Pos	Stemming/ Lemmatization	Accuracy	Time Taken	Classification Report
✓				0.70195	93.4s	precision recall f1-score support 0 0.80 0.53 0.64 9935 1 0.65 0.87 0.75 10065 accuracy 0.70 20000 macro avg 0.73 0.70 0.69 20000 weighted avg 0.73 0.70 0.69 20000
✓	✓			0.70125	131 s	precision recall f1-score support 0 0.80 0.53 0.64 9935 1 0.65 0.87 0.75 10065 accuracy 0.70 20000 macro avg 0.73 0.70 0.69 20000 weighted avg 0.73 0.70 0.69 20000
✓	✓	✓		0.70065	283.6s	precision recall f1-score support 0 0.83 0.50 0.63 9935 1 0.65 0.90 0.75 10065 accuracy 0.70 20000 macro avg 0.74 0.70 0.69 20000 weighted avg 0.74 0.70 0.69 20000

✓	✓		Porter stemmer	0.6571	187.3s	
✓	✓		Lancaster Stemmer	0.68435	163.9s	
✓			Lancaster Stemmer	0.66795	96.72s	
✓	✓		Snowball Stemmer	0.66065	160.9s	
✓	✓		Spacy	0.70485	738.4s	

2- TEXTBLOB

Stop word	Tokenize	Pos	Stemming/ Lemmatization	Accuracy	Time Taken	Classification Report
✓				0.70535	31.3s	
✓	✓			0.7099	75.8 s	

✓	✓	✓		0.60855	374.5s	precision recall f1-score support 0 0.65 0.46 0.54 9935 1 0.59 0.76 0.66 10065 accuracy 0.61 20000 macro avg 0.62 0.61 0.60 20000 weighted avg 0.62 0.61 0.60 20000
✓	✓		Porter stemmer	0.65375	146.8s	precision recall f1-score support 0 0.80 0.40 0.53 9935 1 0.60 0.90 0.72 10065 accuracy 0.65 20000 macro avg 0.70 0.65 0.63 20000 weighted avg 0.70 0.65 0.63 20000
✓	✓		Lancaster Stemmer	0.6676	118.2s	precision recall f1-score support 0 0.75 0.50 0.60 9935 1 0.63 0.83 0.72 10065 accuracy 0.67 20000 macro avg 0.69 0.67 0.66 20000 weighted avg 0.69 0.67 0.66 20000
✓	✓		Snowball Stemmer	0.65515	116.2s	precision recall f1-score support 0 0.81 0.40 0.53 9935 1 0.60 0.91 0.73 10065 accuracy 0.66 20000 macro avg 0.71 0.65 0.63 20000 weighted avg 0.71 0.66 0.63 20000
✓	✓		Spacy	0.7253		Classification Report: precision recall f1-score support negative 0.86 0.53 0.66 9935 positive 0.67 0.91 0.77 10065 accuracy 0.73 20000 macro avg 0.76 0.72 0.71 20000 weighted avg 0.76 0.73 0.71 20000

3- VADER

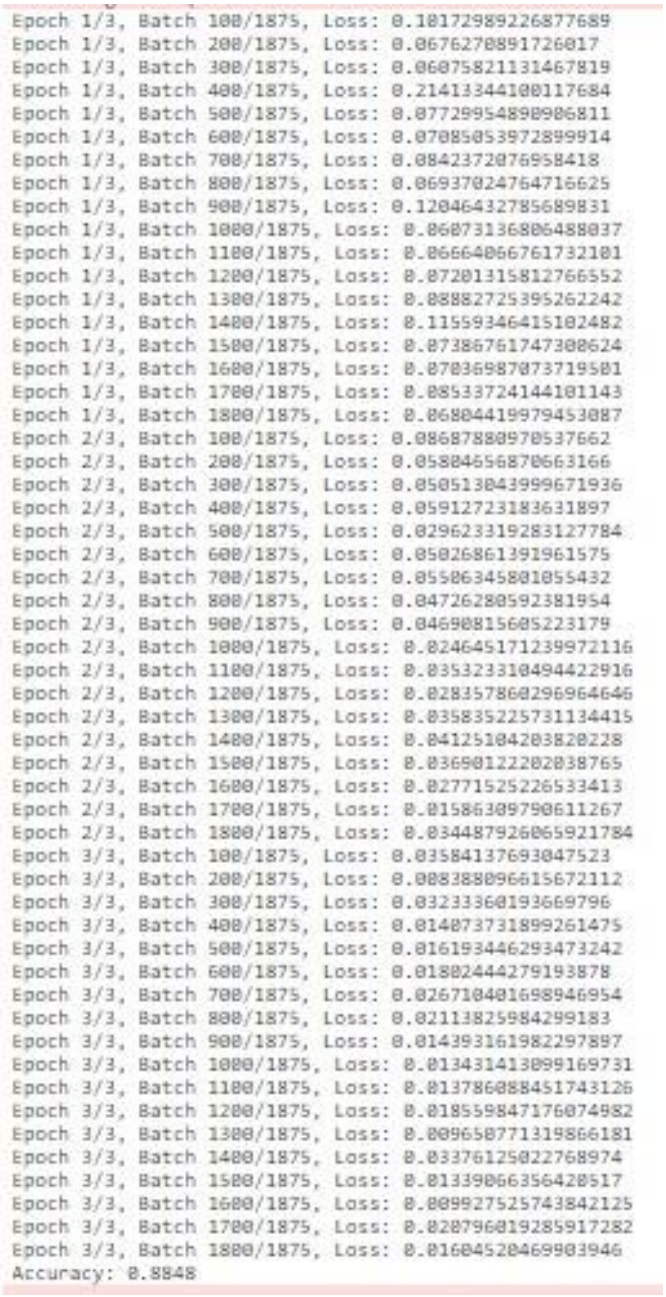
Stop word	Tokenize	Pos	Stemming/ Lemmatization	Accuracy	Time Taken	Classification Report
				0.6998	65.13s	precision recall f1-score support 0 0.79 0.53 0.64 9935 1 0.65 0.86 0.74 10065 accuracy 0.70 20000 macro avg 0.72 0.70 0.69 20000 weighted avg 0.72 0.70 0.69 20000
✓				0.67825	48.84s	precision recall f1-score support 0 0.79 0.48 0.60 9935 1 0.63 0.88 0.73 10065 accuracy 0.68 20000 macro avg 0.71 0.68 0.66 20000 weighted avg 0.71 0.68 0.66 20000

✓	✓			0.68105	89.58s	precision recall f1-score support 0 0.79 0.49 0.60 9935 1 0.63 0.87 0.73 10065 accuracy 0.68 20000 macro avg 0.71 0.68 0.67 20000 weighted avg 0.71 0.68 0.67 20000
✓	✓	✓		0.6763	87.09s	precision recall f1-score support 0 0.79 0.48 0.59 9935 1 0.63 0.87 0.73 10065 accuracy 0.68 20000 macro avg 0.71 0.68 0.66 20000 weighted avg 0.71 0.68 0.66 20000
✓	✓		Porter stemmer	0.67275	207.6s	precision recall f1-score support 0 0.76 0.50 0.61 9935 1 0.63 0.84 0.72 10065 accuracy 0.67 20000 macro avg 0.69 0.67 0.66 20000 weighted avg 0.69 0.67 0.66 20000
✓	✓		Lancaster Stemmer	0.69225	167.5s	precision recall f1-score support 0 0.72 0.63 0.67 9935 1 0.67 0.76 0.71 10065 accuracy 0.69 20000 macro avg 0.70 0.69 0.69 20000 weighted avg 0.70 0.69 0.69 20000
✓	✓		Snowball Stemmer	0.6807	162.8s	precision recall f1-score support 0 0.75 0.53 0.62 9935 1 0.64 0.83 0.72 10065 accuracy 0.68 20000 macro avg 0.70 0.68 0.67 20000 weighted avg 0.70 0.68 0.67 20000
✓	✓		Spacy	0.6762	927.4s	precision recall f1-score support 0 0.78 0.49 0.60 9935 1 0.63 0.86 0.73 10065 accuracy 0.68 20000 macro avg 0.70 0.67 0.66 20000 weighted avg 0.70 0.68 0.66 20000

PLM's

1- Distilbert

Stop Words	Tokenize	Pos	Stemmer/ Lemmatizer	Classification Report & Accuracy
No	No	No	No	max_length = 128 batch_size = 16 scheduler = get_linear_schedule_with_warmup optimizer = AdamW(roberta_model.parameters(), lr=2e-5 Epoch 1/3 - Loss: 0.3393317736208439 Epoch 2/3 - Loss: 0.20182555637632807 Epoch 3/3 - Loss: 0.11509952055377265 Accuracy: 0.85695
No	No	No	No	max_length = 128, batch_size = 16 scaler = GradScaler() num_epochs = 3 max_grad_norm = 1.0

				warmup_steps = 0.1 * len(train_dataloader) scheduler = get_linear_schedule_with_warmup optimizer = AdamW(roberta_model.parameters(), lr=2e-5) 
No	No	No	No	max_length = 128, batch_size = 16 scaler = GradScaler() num_epochs = 3 max_grad_norm = 1.0 warmup_steps = 0.1 * len(train_dataloader) scheduler = get_linear_schedule_with_warmup optimizer = SGD(distilbert_model.parameters(), lr=0.01, momentum=0.9)

				Epoch 1/3, Batch 100/1875, Loss: 0.638399543762207 Epoch 1/3, Batch 200/1875, Loss: 0.494405632019043 Epoch 1/3, Batch 300/1875, Loss: 0.422783145904541 Epoch 1/3, Batch 400/1875, Loss: 0.4123725938796997 Epoch 1/3, Batch 500/1875, Loss: 0.3828009366989136 Epoch 1/3, Batch 600/1875, Loss: 0.380307683944702 16 Epoch 1/3, Batch 700/1875, Loss: 0.378375313282812 96 Epoch 1/3, Batch 800/1875, Loss: 0.4078387808799744 Epoch 1/3, Batch 900/1875, Loss: 0.3658698034286499 Epoch 1/3, Batch 1000/1875, Loss: 0.3521576488018036 Epoch 1/3, Batch 1100/1875, Loss: 0.344458261592865 Epoch 1/3, Batch 1200/1875, Loss: 0.3489921288490295 Epoch 1/3, Batch 1300/1875, Loss: 0.3452486753463745 Epoch 1/3, Batch 1400/1875, Loss: 0.360827047824859 64 Epoch 1/3, Batch 1500/1875, Loss: 0.320198940038681 05 Epoch 1/3, Batch 1600/1875, Loss: 0.332977634668350 23 Epoch 1/3, Batch 1700/1875, Loss: 0.362967230081558 23 Epoch 1/3, Batch 1800/1875, Loss: 0.3260021793842316 Epoch 2/3, Batch 100/1875, Loss: 0.329856047630310 07 Epoch 2/3, Batch 200/1875, Loss: 0.315281245708465 55 Epoch 2/3, Batch 300/1875, Loss: 0.28176306784152 98 Epoch 2/3, Batch 400/1875, Loss: 0.301150300502777 1 Epoch 2/3, Batch 500/1875, Loss: 0.280997711122036 Epoch 2/3, Batch 600/1875, Loss: 0.286783331632614 16 Epoch 2/3, Batch 700/1875, Loss: 0.25768566429615 02 Epoch 2/3, Batch 800/1875, Loss: 0.267879300880432 14 Epoch 2/3, Batch 900/1875, Loss: 0.25697315633296 97 Epoch 2/3, Batch 1000/1875, Loss: 0.2364058029651642 Epoch 2/3, Batch 1100/1875, Loss: 0.233076520860195 15 Epoch 2/3, Batch 1200/1875, Loss: 0.233382978439331 06 Epoch 2/3, Batch 1300/1875, Loss: 0.27006321966648 1 Epoch 2/3, Batch 1400/1875, Loss: 0.27290216207504 27 Epoch 2/3, Batch 1500/1875, Loss: 0.221527493000030 52 Epoch 2/3, Batch 1600/1875, Loss: 0.272287574410438 55 Epoch 2/3, Batch 1700/1875, Loss: 0.250266785621643 86 Epoch 2/3, Batch 1800/1875, Loss: 0.220980866555399 23 Epoch 3/3, Batch 100/1875, Loss: 0.24657639414072036 Epoch 3/3, Batch 200/1875, Loss: 0.2323412564396858 Epoch 3/3, Batch 300/1875, Loss: 0.19998804360628128 Epoch 3/3, Batch 400/1875, Loss: 0.21420256346464156 Epoch 3/3, Batch 500/1875, Loss: 0.18530170395970344 Epoch 3/3, Batch 600/1875, Loss: 0.19853367209434508 Epoch 3/3, Batch 700/1875, Loss: 0.1722753517329693 Epoch 3/3, Batch 800/1875, Loss: 0.19756732687354087 Epoch 3/3, Batch 900/1875, Loss: 0.18920622900128364 Epoch 3/3, Batch 1000/1875, Loss: 0.16169812321662902 Epoch 3/3, Batch 1100/1875, Loss: 0.1534866251051426 Epoch 3/3, Batch 1200/1875, Loss: 0.158273472048891 65 Epoch 3/3, Batch 1300/1875, Loss: 0.167168603688478 47 Epoch 3/3, Batch 1400/1875, Loss: 0.21860954046249 39 Epoch 3/3, Batch 1500/1875, Loss: 0.18141757585108 28 Epoch 3/3, Batch 1600/1875, Loss: 0.20031207814812 66 Epoch 3/3, Batch 1700/1875, Loss: 0.189339234915028572 Epoch 3/3, Batch 1800/1875, Loss: 0.17760580018162728 Accuracy: 0.87725
No	No	No	No	max_length = 128, batch_size = 16 scaler = GradScaler() num_epochs = 3 max_grad_norm = 1.0 warmup_steps = 0.1 * len(train_dataloader) optimizer = AdamW(roberta_model.parameters(), lr=2e-5 scheduler = ExponentialLR(optimizer_roberta, gamma=0.9) Epoch 1/3, Batch 100/1875, Loss: 0.6952365112304687 Epoch 1/3, Batch 200/1875, Loss: 0.696234130859375 Epoch 1/3, Batch 300/1875, Loss: 0.6951004028320312 Epoch 1/3, Batch 400/1875, Loss: 0.6953091430664062 Epoch 1/3, Batch 500/1875, Loss: 0.6928515625 Epoch 1/3, Batch 600/1875, Loss: 0.694931640625 Epoch 1/3, Batch 700/1875, Loss: 0.696868115234375 Epoch 1/3, Batch 800/1875, Loss: 0.695142822265625 Epoch 1/3, Batch 900/1875, Loss: 0.6942910766601562 Epoch 1/3, Batch 1000/1875, Loss: 0.69353759765625 Epoch 1/3, Batch 1100/1875, Loss: 0.6954696655273438 Epoch 1/3, Batch 1200/1875, Loss: 0.6924240112304687 Epoch 1/3, Batch 1300/1875, Loss: 0.6942092895507812 Epoch 1/3, Batch 1400/1875, Loss: 0.69363525390625 Epoch 1/3, Batch 1500/1875, Loss: 0.6938516235351563 Epoch 1/3, Batch 1600/1875, Loss: 0.694998779296875 Epoch 1/3, Batch 1700/1875, Loss: 0.6942047119140625 Epoch 1/3, Batch 1800/1875, Loss: 0.6917852783203124 Epoch 2/3, Batch 100/1875, Loss: 0.695845947265625 Epoch 2/3, Batch 200/1875, Loss: 0.6955203247070313 Epoch 2/3, Batch 300/1875, Loss: 0.6943658447265625 Epoch 2/3, Batch 400/1875, Loss: 0.6958830078125 Epoch 2/3, Batch 500/1875, Loss: 0.692049560546875 Epoch 2/3, Batch 600/1875, Loss: 0.6933535766601563 Epoch 2/3, Batch 700/1875, Loss: 0.6917349243164063 Epoch 2/3, Batch 800/1875, Loss: 0.6946011352539062 Epoch 2/3, Batch 900/1875, Loss: 0.69296875 Epoch 2/3, Batch 1000/1875, Loss: 0.6924227905273438 Epoch 2/3, Batch 1100/1875, Loss: 0.6914770507812 5

				Epoch 2/3, Batch 1000/1875, Loss: 0.6924227905273438 Epoch 2/3, Batch 1100/1875, Loss: 0.69147705078125 Epoch 2/3, Batch 1200/1875, Loss: 0.6934478759765625 Epoch 2/3, Batch 1300/1875, Loss: 0.6943014526367187 Epoch 2/3, Batch 1400/1875, Loss: 0.695115966796875 Epoch 2/3, Batch 1500/1875, Loss: 0.6943798828125 Epoch 2/3, Batch 1600/1875, Loss: 0.6947116088867188 Epoch 2/3, Batch 1700/1875, Loss: 0.693489990234375 Epoch 2/3, Batch 1800/1875, Loss: 0.6922604370117188 Epoch 3/3, Batch 100/1875, Loss: 0.6951278686523438 Epoch 3/3, Batch 200/1875, Loss: 0.6953207397460938 Epoch 3/3, Batch 300/1875, Loss: 0.69247802734375 Epoch 3/3, Batch 400/1875, Loss: 0.695355224609375 Epoch 3/3, Batch 500/1875, Loss: 0.6948406982421875 Epoch 3/3, Batch 600/1875, Loss: 0.6944290161132812 Epoch 3/3, Batch 700/1875, Loss: 0.6951596069335938 Epoch 3/3, Batch 800/1875, Loss: 0.6949166870117187 Epoch 3/3, Batch 900/1875, Loss: 0.6925833129882812 Epoch 3/3, Batch 1000/1875, Loss: 0.6930105590820312 Epoch 3/3, Batch 1100/1875, Loss: 0.6935324096679687 Epoch 3/3, Batch 1200/1875, Loss: 0.6939352416992187 Epoch 3/3, Batch 1300/1875, Loss: 0.6931344604492188 Epoch 3/3, Batch 1400/1875, Loss: 0.6958328247070312 Epoch 3/3, Batch 1500/1875, Loss: 0.6924664306640625 Epoch 3/3, Batch 1600/1875, Loss: 0.6954547119140625 Epoch 3/3, Batch 1700/1875, Loss: 0.6957962036132812 Epoch 3/3, Batch 1800/1875, Loss: 0.6926739501953125 Accuracy: 0.5028
No	No	No	No	Finetuning Accuracy: 0.93185 F1-score: 0.9313349599163471 Confusion matrix: [[9269 666] [713 9352]] precision recall f1-score support negative 0.93 0.93 0.93 9935 positive 0.93 0.93 0.93 10065 accuracy 0.93 20000 macro avg 0.93 0.93 0.93 20000 weighted avg 0.93 0.93 0.93 20000 Time required to fine-tune: 3672.2959537506104

2- Roberta

Stop Words	Tokenize	Pos	Stemmer/ Lemmatizer	Classification Report & Accuracy
No	No	No	No	max_length = 128, batch_size = 16 optimizer = AdamW(roberta_model.parameters(), lr=2e-5 scheduler = get_linear_schedule_with_warmup Epoch 1/3 - Loss: 0.6953605392773946 Epoch 2/3 - Loss: 0.6947482631047567 Epoch 3/3 - Loss: 0.6946594745000203 Accuracy: 0.49675
				max_length = 128, batch_size = 16 scaler = GradScaler(), num_epochs = 3, max_grad_norm = 1.0 warmup_steps = 0.1 * len(train_dataloader) optimizer = AdamW(roberta_model.parameters(), lr=2e-5 scheduler = get_linear_schedule_with_warmup Epoch 1/3, Batch 100/1875, Loss: 0.6939346313476562 Epoch 1/3, Batch 200/1875, Loss: 0.7001318359375 Epoch 1/3, Batch 300/1875, Loss: 0.6976708984375 Epoch 1/3, Batch 400/1875, Loss: 0.6888579559326172 Epoch 1/3, Batch 500/1875, Loss: 0.6961116790771484 Epoch 1/3, Batch 600/1875, Loss: 0.6953240966796875 Epoch 1/3, Batch 700/1875, Loss: 0.6958424377441407 Epoch 1/3, Batch 800/1875, Loss: 0.6959896850585937 Epoch 1/3, Batch 900/1875, Loss: 0.6935813903808594 Epoch 1/3, Batch 1000/1875, Loss: 0.6943628300292969 Epoch 1/3, Batch 1100/1875, Loss: 0.6938188171386719 Epoch 1/3, Batch 1200/1875, Loss: 0.6974050903320312 Epoch 1/3, Batch 1300/1875, Loss: 0.6944967651367188 Epoch 1/3, Batch 1400/1875, Loss: 0.6878280639648438 Epoch 1/3, Batch 1500/1875, Loss: 0.6747649765014648 Epoch 1/3, Batch 1600/1875, Loss: 0.6593856048583985 Epoch 1/3, Batch 1700/1875, Loss: 0.6286680984497071 Epoch 1/3, Batch 1800/1875, Loss: 0.5979949188232422 Epoch 2/3, Batch 100/1875, Loss: 0.5938976287841797 Epoch 2/3, Batch 200/1875, Loss: 0.5936203193664551 Epoch 2/3, Batch 300/1875, Loss: 0.5665141296386719 Epoch 2/3, Batch 400/1875, Loss: 0.5586728858947754 Epoch 2/3, Batch 500/1875, Loss: 0.5395234870910645 Epoch 2/3, Batch 600/1875, Loss: 0.5487950325012207 Epoch 2/3, Batch 700/1875, Loss: 0.5115027236938476 Epoch 2/3, Batch 800/1875, Loss: 0.5339586687088013 Epoch 2/3, Batch 900/1875, Loss: 0.5137767028808594

				Epoch 2/3, Batch 900/1875, Loss: 0.5137767028808594 Epoch 2/3, Batch 1000/1875, Loss: 0.5034670782089233 Epoch 2/3, Batch 1100/1875, Loss: 0.483096604347229 Epoch 2/3, Batch 1200/1875, Loss: 0.5047172737121582 Epoch 2/3, Batch 1300/1875, Loss: 0.47947437286376954 Epoch 2/3, Batch 1400/1875, Loss: 0.4923399066925049 Epoch 2/3, Batch 1500/1875, Loss: 0.46708165168762206 Epoch 2/3, Batch 1600/1875, Loss: 0.4586261510848999 Epoch 2/3, Batch 1700/1875, Loss: 0.4527506160736084 Epoch 2/3, Batch 1800/1875, Loss: 0.4514769458770752 Epoch 3/3, Batch 100/1875, Loss: 0.4637337350845337 Epoch 3/3, Batch 200/1875, Loss: 0.4725750160217285 Epoch 3/3, Batch 300/1875, Loss: 0.4398309047698975 Epoch 3/3, Batch 400/1875, Loss: 0.4431319189071655 Epoch 3/3, Batch 500/1875, Loss: 0.44276777029037473 Epoch 3/3, Batch 600/1875, Loss: 0.4567181968688965 Epoch 3/3, Batch 700/1875, Loss: 0.4231038761138916 Epoch 3/3, Batch 800/1875, Loss: 0.43161486148834227 Epoch 3/3, Batch 900/1875, Loss: 0.427634220123291 Epoch 3/3, Batch 1000/1875, Loss: 0.428695821762085 Epoch 3/3, Batch 1100/1875, Loss: 0.4082297658920288 Epoch 3/3, Batch 1200/1875, Loss: 0.42948600292205813 Epoch 3/3, Batch 1300/1875, Loss: 0.42375107765197756 Epoch 3/3, Batch 1400/1875, Loss: 0.43669298648834226 Epoch 3/3, Batch 1500/1875, Loss: 0.3974007177352905 Epoch 3/3, Batch 1600/1875, Loss: 0.3981157207489014 Epoch 3/3, Batch 1700/1875, Loss: 0.39326844453811644 Epoch 3/3, Batch 1800/1875, Loss: 0.3862108492851257 Accuracy: 0.7854
No	No	No	No	max_length = 128, batch_size = 16 scaler = GradScaler() num_epochs = 3 max_grad_norm = 1.0 warmup_steps = 0.1 * len(train_dataloader) optimizer = SGD(distilbert_model.parameters(), lr=0.01, momentum=0.9) scheduler = get_linear_schedule_with_warmup Epoch 1/3, Batch 100/1875, Loss: 0.7025511932373046 Epoch 1/3, Batch 200/1875, Loss: 0.7114691925048828 Epoch 1/3, Batch 300/1875, Loss: 0.7070757293701172 Epoch 1/3, Batch 400/1875, Loss: 0.7067382287353516 Epoch 1/3, Batch 500/1875, Loss: 0.6982167053222657 Epoch 1/3, Batch 600/1875, Loss: 0.6987605285644531 Epoch 1/3, Batch 700/1875, Loss: 0.6948300170898437 Epoch 1/3, Batch 800/1875, Loss: 0.69453857421875 Epoch 1/3, Batch 900/1875, Loss: 0.69548095703125 Epoch 1/3, Batch 1000/1875, Loss: 0.6946369934082032 Epoch 1/3, Batch 1100/1875, Loss: 0.6942010498046876 Epoch 1/3, Batch 1200/1875, Loss: 0.6946176147460937 Epoch 1/3, Batch 1300/1875, Loss: 0.6954307556152344 Epoch 1/3, Batch 1400/1875, Loss: 0.694906005859375 Epoch 1/3, Batch 1500/1875, Loss: 0.6949349975585938 Epoch 1/3, Batch 1600/1875, Loss: 0.6954974365234375 Epoch 1/3, Batch 1700/1875, Loss: 0.69441162109375 Epoch 1/3, Batch 1800/1875, Loss: 0.69405029296875 Epoch 2/3, Batch 100/1875, Loss: 0.694393310546875 Epoch 2/3, Batch 200/1875, Loss: 0.6943914794921875 Epoch 2/3, Batch 300/1875, Loss: 0.693854980046875 Epoch 2/3, Batch 400/1875, Loss: 0.6942315673828126 Epoch 2/3, Batch 500/1875, Loss: 0.6952951049804688 Epoch 2/3, Batch 600/1875, Loss: 0.6945474243164063 Epoch 2/3, Batch 700/1875, Loss: 0.6922427368164062 Epoch 2/3, Batch 800/1875, Loss: 0.6928944396972656 Epoch 2/3, Batch 900/1875, Loss: 0.6942579650878906 Epoch 2/3, Batch 1000/1875, Loss: 0.6932888793945312 Epoch 2/3, Batch 1100/1875, Loss: 0.6932589721679687 Epoch 2/3, Batch 1200/1875, Loss: 0.6944659423828125 Epoch 2/3, Batch 1300/1875, Loss: 0.6943551635742188 Epoch 2/3, Batch 1400/1875, Loss: 0.6934896850585938 Epoch 2/3, Batch 1500/1875, Loss: 0.6932945251464844 Epoch 2/3, Batch 1600/1875, Loss: 0.693909912109375 Epoch 2/3, Batch 1700/1875, Loss: 0.6935284423828125 Epoch 2/3, Batch 1800/1875, Loss: 0.6941763305664063 Epoch 3/3, Batch 100/1875, Loss: 0.693385925292968 Epoch 3/3, Batch 200/1875, Loss: 0.6931289672851563 Epoch 3/3, Batch 300/1875, Loss: 0.693884716796875 Epoch 3/3, Batch 400/1875, Loss: 0.6935757446289063 Epoch 3/3, Batch 500/1875, Loss: 0.6938406372070313 Epoch 3/3, Batch 600/1875, Loss: 0.6936569213867188 Epoch 3/3, Batch 700/1875, Loss: 0.6927560424804687 Epoch 3/3, Batch 800/1875, Loss: 0.6925009155273437 Epoch 3/3, Batch 600/1875, Loss: 0.6936569213867188 Epoch 3/3, Batch 700/1875, Loss: 0.6927560424804687 Epoch 3/3, Batch 800/1875, Loss: 0.6925009155273437 Epoch 3/3, Batch 900/1875, Loss: 0.6931149291992188 Epoch 3/3, Batch 1000/1875, Loss: 0.692882080078125 Epoch 3/3, Batch 1100/1875, Loss: 0.6899261474609375 Epoch 3/3, Batch 1200/1875, Loss: 0.68900390625 Epoch 3/3, Batch 1300/1875, Loss: 0.6943078613281125 Epoch 3/3, Batch 1400/1875, Loss: 0.6934947204589844 Epoch 3/3, Batch 1500/1875, Loss: 0.6876557922363281 Epoch 3/3, Batch 1600/1875, Loss: 0.6826687622070312 Epoch 3/3, Batch 1700/1875, Loss: 0.6789578247070313 Epoch 3/3, Batch 1800/1875, Loss: 0.678226318359375 Accuracy: 0.5969
No	No	No	No	max_length = 128,

				batch_size = 16 scaler = GradScaler() num_epochs = 3 max_grad_norm = 1.0 warmup_steps = 0.1 * len(train_dataloader) optimizer = AdamW(roberta_model.parameters(), lr=2e-5 scheduler = ExponentialLR(optimizer_roberta, gamma=0.9) Epoch 1/3, Batch 100/1875, Loss: 0.7008372497558594 Epoch 1/3, Batch 200/1875, Loss: 0.6993682861328125 Epoch 1/3, Batch 300/1875, Loss: 0.6985882568359375 Epoch 1/3, Batch 400/1875, Loss: 0.7031089782714843 Epoch 1/3, Batch 500/1875, Loss: 0.6975228881835938 Epoch 1/3, Batch 600/1875, Loss: 0.6984939575195312 Epoch 1/3, Batch 700/1875, Loss: 0.7010128784179688 Epoch 1/3, Batch 800/1875, Loss: 0.6990983581542969 Epoch 1/3, Batch 900/1875, Loss: 0.692613525390625 Epoch 1/3, Batch 1000/1875, Loss: 0.6938954162597656 Epoch 1/3, Batch 1100/1875, Loss: 0.6946134948730469 Epoch 1/3, Batch 1200/1875, Loss: 0.6988857849121094 Epoch 1/3, Batch 1300/1875, Loss: 0.6969267272949219 Epoch 1/3, Batch 1400/1875, Loss: 0.6992776489257813 Epoch 1/3, Batch 1500/1875, Loss: 0.6999372863769531 Epoch 1/3, Batch 1600/1875, Loss: 0.6993745422363281 Epoch 1/3, Batch 1700/1875, Loss: 0.6991278076171875 Epoch 1/3, Batch 1800/1875, Loss: 0.6999179077148437 Epoch 2/3, Batch 100/1875, Loss: 0.7030986022949218 Epoch 2/3, Batch 200/1875, Loss: 0.7003971862792969 Epoch 2/3, Batch 300/1875, Loss: 0.698895263671875 Epoch 2/3, Batch 400/1875, Loss: 0.7002685546875 Epoch 2/3, Batch 500/1875, Loss: 0.6960035705566406 Epoch 2/3, Batch 600/1875, Loss: 0.7009284973144532 Epoch 2/3, Batch 700/1875, Loss: 0.6981956481933593 Epoch 2/3, Batch 800/1875, Loss: 0.699626770019531 Epoch 2/3, Batch 900/1875, Loss: 0.6964289855957031 Epoch 2/3, Batch 1000/1875, Loss: 0.6953314208984375 Epoch 2/3, Batch 1100/1875, Loss: 0.6953016662597656 Epoch 2/3, Batch 1200/1875, Loss: 0.6975892639160156 Epoch 2/3, Batch 1300/1875, Loss: 0.6971302795410156 Epoch 2/3, Batch 1400/1875, Loss: 0.6975001525878907 Epoch 2/3, Batch 1500/1875, Loss: 0.6958174133300781 Epoch 2/3, Batch 1600/1875, Loss: 0.6976278686523437 Epoch 2/3, Batch 1700/1875, Loss: 0.6980998229980468 Epoch 2/3, Batch 1800/1875, Loss: 0.6967844067382813 Epoch 3/3, Batch 100/1875, Loss: 0.7012713623046875 Epoch 3/3, Batch 200/1875, Loss: 0.6991055297851563 Epoch 3/3, Batch 300/1875, Loss: 0.6986656188964844 Epoch 3/3, Batch 400/1875, Loss: 0.6993418884277344 Epoch 3/3, Batch 500/1875, Loss: 0.6976358032226563 Epoch 3/3, Batch 600/1875, Loss: 0.6989553833007812 Epoch 3/3, Batch 700/1875, Loss: 0.7009678649902343 Epoch 3/3, Batch 800/1875, Loss: 0.6979411315917968 Epoch 3/3, Batch 900/1875, Loss: 0.6952255249023438 Epoch 3/3, Batch 1000/1875, Loss: 0.6954566955566406 Epoch 3/3, Batch 1100/1875, Loss: 0.6937945556640625 Epoch 3/3, Batch 1200/1875, Loss: 0.695462646484375 Epoch 3/3, Batch 1300/1875, Loss: 0.6949417114257812 Epoch 3/3, Batch 1400/1875, Loss: 0.6965745544433594 Epoch 3/3, Batch 1500/1875, Loss: 0.6979829406738282 Epoch 3/3, Batch 1600/1875, Loss: 0.6994242858886719 Epoch 3/3, Batch 1700/1875, Loss: 0.6975149536132812 Epoch 3/3, Batch 1800/1875, Loss: 0.6957183837890625 Accuracy: 0.50325
No	No	No	No	Finetuning Accuracy: 0.95255 F1-score: 0.9531149646756584 Confusion matrix: [[9485 530] [419 9646]] precision recall f1-score support negative 0.96 0.95 0.95 9935 positive 0.95 0.96 0.95 10065 accuracy 0.95 20000 macro avg 0.95 0.95 0.95 20000 weighted avg 0.95 0.95 0.95 20000 Time required to fine-tune: 6715.162095785141

Word Embedding:

1- Word2Vec

Stop Words	Tokenize	Pos	Stemmer/ Lemmatizer	Classification Report & Accuracy
No	No	No	No	Decision Tree Accuracy: 0.6451 Decision Tree Classification Report: precision recall f1-score support positive 0.65 0.63 0.64 9935 negative 0.64 0.66 0.65 10065 accuracy 0.65 20000 macro avg 0.65 0.65 0.64 20000 weighted avg 0.65 0.65 0.65 20000 Random Forest Accuracy: 0.75375 Random Forest Classification Report: precision recall f1-score support positive 0.76 0.74 0.75 9935 negative 0.75 0.77 0.76 10065 accuracy 0.75 20000 macro avg 0.75 0.75 0.75 20000 weighted avg 0.75 0.75 0.75 20000 Logistic Regression Accuracy: 0.80365 Logistic Regression Classification Report: precision recall f1-score support positive 0.81 0.79 0.80 9935 negative 0.80 0.82 0.81 10065 accuracy 0.80 20000 macro avg 0.80 0.80 0.80 20000 weighted avg 0.80 0.80 0.80 20000 k-NN Accuracy: 0.7025 k-NN Classification Report: precision recall f1-score support positive 0.69 0.73 0.71 9935 negative 0.72 0.68 0.70 10065 accuracy 0.70 20000 macro avg 0.70 0.70 0.70 20000 weighted avg 0.70 0.70 0.70 20000 XGBoost Accuracy: 0.77965 XGBoost Classification Report: precision recall f1-score support positive 0.79 0.77 0.78 9935 negative 0.77 0.79 0.78 10065 accuracy 0.78 20000 macro avg 0.78 0.78 0.78 20000 weighted avg 0.78 0.78 0.78 20000

				LightGBM Accuracy: 0.78095 LightGBM Classification Report: precision recall f1-score support positive 0.79 0.77 0.78 9935 negative 0.77 0.80 0.79 10065 accuracy 0.78 20000 macro avg 0.78 0.78 0.78 20000 weighted avg 0.78 0.78 0.78 20000 Gradient Boosting Accuracy: 0.76545 Gradient Boosting Classification Report: precision recall f1-score support positive 0.77 0.75 0.76 9935 negative 0.76 0.78 0.77 10065 accuracy 0.77 20000 macro avg 0.77 0.77 0.77 20000 weighted avg 0.77 0.77 0.77 20000 Neural Network Accuracy: 0.80475 Neural Network Classification Report: precision recall f1-score support positive 0.82 0.78 0.80 9935 negative 0.79 0.83 0.81 10065 accuracy 0.80 20000 macro avg 0.81 0.80 0.80 20000 weighted avg 0.81 0.80 0.80 20000
No	No	No	No	Hyper Parameter Tuning Decision Tree Accuracy: 0.66285 Decision Tree Classification Report: precision recall f1-score support positive 0.66 0.67 0.66 9935 negative 0.67 0.66 0.66 10065 accuracy 0.66 20000 macro avg 0.66 0.66 0.66 20000 weighted avg 0.66 0.66 0.66 20000 Random Forest Accuracy: 0.76105 Random Forest Classification Report: precision recall f1-score support positive 0.77 0.73 0.75 9935 negative 0.75 0.79 0.77 10065 accuracy 0.76 20000 macro avg 0.76 0.76 0.76 20000 weighted avg 0.76 0.76 0.76 20000 Logistic Regression Accuracy: 0.80655 Logistic Regression Classification Report: precision recall f1-score support positive 0.81 0.79 0.80 9935 negative 0.80 0.82 0.81 10065 accuracy 0.81 20000 macro avg 0.81 0.81 0.81 20000 weighted avg 0.81 0.81 0.81 20000

				LightGBM Accuracy: 0.7816 LightGBM Classification Report: precision recall f1-score support positive 0.79 0.76 0.77 9935 negative 0.77 0.81 0.79 10065 accuracy 0.78 20000 macro avg 0.78 0.78 0.78 20000 weighted avg 0.78 0.78 0.78 20000 Gradient Boosting Accuracy: 0.7912 Gradient Boosting Classification Report: precision recall f1-score support positive 0.80 0.77 0.79 9935 negative 0.78 0.81 0.80 10065 accuracy 0.79 20000 macro avg 0.79 0.79 0.79 20000 weighted avg 0.79 0.79 0.79 20000 Neural Network Accuracy: 0.80535 Neural Network Classification Report: precision recall f1-score support positive 0.79 0.83 0.81 9935 negative 0.83 0.78 0.80 10065 accuracy 0.81 20000 macro avg 0.81 0.81 0.81 20000 weighted avg 0.81 0.81 0.81 20000 k-NN Accuracy: 0.7159 k-NN Classification Report: precision recall f1-score support positive 0.70 0.75 0.72 9935 negative 0.73 0.68 0.71 10065 accuracy 0.72 20000 macro avg 0.72 0.72 0.72 20000 weighted avg 0.72 0.72 0.72 20000 XGBoost Accuracy: 0.7923 XGBoost Classification Report: precision recall f1-score support positive 0.80 0.77 0.79 9935 negative 0.78 0.81 0.80 10065 accuracy 0.79 20000 macro avg 0.79 0.79 0.79 20000 weighted avg 0.79 0.79 0.79 20000
No	No	No	No	Parameters for the model and training max_features = 20000 maxlen = 200 embedding_dim = 300 lstm_units = 128 filters = 64 kernel_size = 5 batch_size = 32 epochs = 5 lstm_branch = LSTM(lstm_units, dropout=0.2, recurrent_dropout=0.2)(embedding_layer) cnn_branch = Conv1D(filters, kernel_size, activation='relu')(embedding_layer) cnn_branch = GlobalMaxPooling1D()(cnn_branch) merged = Concatenate()([lstm_branch, cnn_branch]) output = Dense(1, activation='sigmoid')(merged)

				Epoch 1/5 750/750 [==========] - 187s 247ms/step - loss: 0.3793 - accuracy: 0.8282 - val_loss: 0.3028 - val_accuracy: 0.8757 Epoch 2/5 750/750 [==========] - 183s 244ms/step - loss: 0.2624 - accuracy: 0.8917 - val_loss: 0.2874 - val_accuracy: 0.8827 Epoch 3/5 750/750 [==========] - 184s 246ms/step - loss: 0.2004 - accuracy: 0.9241 - val_loss: 0.2600 - val_accuracy: 0.8970 Epoch 4/5 750/750 [==========] - 184s 245ms/step - loss: 0.1470 - accuracy: 0.9506 - val_loss: 0.2584 - val_accuracy: 0.8968 Epoch 5/5 750/750 [==========] - 183s 244ms/step - loss: 0.1005 - accuracy: 0.9722 - val_loss: 0.2737 - val_accuracy: 0.8958 625/625 [==========] - 32s 51ms/step - loss: 0.2810 - accuracy: 0.8896 Test accuracy: 0.8895999789237976 Time required to fine-tune: 1033.9831807613373
No	No	No	No	lstm_branch = LSTM(lstm_units, dropout=0.2, recurrent_dropout=0.2)(embedding_layer) cnn_branch = Conv1D(filters, kernel_size, activation='relu')(embedding_layer) cnn_branch = GlobalMaxPooling1D()(cnn_branch) merged = Concatenate()([lstm_branch, cnn_branch]) merged = Dropout(0.5)(merged) output = Dense(1, activation='sigmoid')(merged) Epoch 1/5 750/750 [==========] - 193s 254ms/step - loss: 0.4272 - accuracy: 0.7999 - val_loss: 0.3195 - val_accuracy: 0.8647 Epoch 2/5 750/750 [==========] - 189s 252ms/step - loss: 0.3324 - accuracy: 0.8567 - val_loss: 0.3146 - val_accuracy: 0.8707 Epoch 3/5 750/750 [==========] - 187s 249ms/step - loss: 0.2883 - accuracy: 0.8793 - val_loss: 0.2803 - val_accuracy: 0.8852 Epoch 4/5 750/750 [==========] - 186s 248ms/step - loss: 0.2568 - accuracy: 0.8934 - val_loss: 0.3061 - val_accuracy: 0.8720 Epoch 5/5 750/750 [==========] - 187s 249ms/step - loss: 0.2285 - accuracy: 0.9068 - val_loss: 0.2695 - val_accuracy: 0.8873 625/625 [==========] - 33s 52ms/step - loss: 0.2660 - accuracy: 0.8881 Test accuracy: 0.8881000280380249 Time required to fine-tune: 1017.9823062419891
No	No	No	No	max_features = 20000 maxlen = 200 embedding_dim = 300 lstm_units = 128 filters = 64 kernel_size = 5 batch_size = 32 epochs = 10 Epoch 1/10 750/750 [==========] - 192s 253ms/step - loss: 0.3904 - accuracy: 0.8227 - val_loss: 0.3287 - val_accuracy: 0.8582 Epoch 2/10 750/750 [==========] - 190s 254ms/step - loss: 0.2673 - accuracy: 0.8880 - val_loss: 0.2803 - val_accuracy: 0.8843 Epoch 3/10 750/750 [==========] - 189s 252ms/step - loss: 0.2069 - accuracy: 0.9198 - val_loss: 0.2781 - val_accuracy: 0.8878 Epoch 4/10 750/750 [==========] - 188s 250ms/step - loss: 0.1532 - accuracy: 0.9469 - val_loss: 0.2635 - val_accuracy: 0.8962 Epoch 5/10 750/750 [==========] - 187s 249ms/step - loss: 0.1071 - accuracy: 0.9674 - val_loss: 0.2753 - val_accuracy: 0.8920 Epoch 6/10 750/750 [==========] - 188s 250ms/step - loss: 0.0709 - accuracy: 0.9846 - val_loss: 0.2793 - val_accuracy: 0.8992 Epoch 7/10 750/750 [==========] - 188s 251ms/step - loss: 0.0429 - accuracy: 0.9951 - val_loss: 0.2943 - val_accuracy: 0.8980 Epoch 8/10 750/750 [==========] - 189s 252ms/step - loss: 0.0256 - accuracy: 0.9984 - val_loss: 0.3196 - val_accuracy: 0.8977 Epoch 9/10 750/750 [==========] - 188s 250ms/step - loss: 0.0155 - accuracy: 0.9996 - val_loss: 0.3314 - val_accuracy: 0.8982 Epoch 10/10 750/750 [==========] - 189s 253ms/step - loss: 0.0092 - accuracy: 0.9999 - val_loss: 0.3569 - val_accuracy: 0.8975 625/625 [==========] - 33s 53ms/step - loss: 0.3658 - accuracy: 0.8899 Test accuracy: 0.8899499773979187 Time required to fine-tune: 2024.7742836475372
No	No	No	Porter Stemmer	[nltk_data] Downloading package punkt to /usr/share/nltk_data... [nltk_data] Package punkt is already up-to-date! Epoch 1/5 750/750 [==========] - 201s 265ms/step - loss: 0.4524 - accuracy: 0.7826 - val_loss: 0.3905 - val_accuracy: 0.8123 Epoch 2/5 750/750 [==========] - 191s 254ms/step - loss: 0.3149 - accuracy: 0.8660 - val_loss: 0.3301 - val_accuracy: 0.8513 Epoch 3/5 750/750 [==========] - 193s 257ms/step - loss: 0.2432 - accuracy: 0.9041 - val_loss: 0.3113 - val_accuracy: 0.8597 Epoch 4/5 750/750 [==========] - 200s 266ms/step - loss: 0.1839 - accuracy: 0.9350 - val_loss: 0.3109 - val_accuracy: 0.8652 Epoch 5/5 750/750 [==========] - 190s 254ms/step - loss: 0.1325 - accuracy: 0.9599 - val_loss: 0.3415 - val_accuracy: 0.8590 625/625 [==========] - 33s 53ms/step - loss: 0.3468 - accuracy: 0.8576 Test accuracy: 0.8576499819755554 Time required to fine-tune: 1375.0074853897095
No	No	No	Lancaster Stemmer	[nltk_data] Downloading package punkt to /usr/share/nltk_data... [nltk_data] Package punkt is already up-to-date! Epoch 1/5 750/750 [==========] - 260s 343ms/step - loss: 0.4463 - accuracy: 0.7897 - val_loss: 0.3717 - val_accuracy: 0.8320 Epoch 2/5 750/750 [==========] - 258s 343ms/step - loss: 0.3102 - accuracy: 0.8677 - val_loss: 0.3265 - val_accuracy: 0.8570 Epoch 3/5 750/750 [==========] - 257s 343ms/step - loss: 0.2358 - accuracy: 0.9081 - val_loss: 0.3074 - val_accuracy: 0.8690 Epoch 4/5 750/750 [==========] - 260s 346ms/step - loss: 0.1791 - accuracy: 0.9364 - val_loss: 0.3025 - val_accuracy: 0.8743 Epoch 5/5 750/750 [==========] - 262s 349ms/step - loss: 0.1274 - accuracy: 0.9612 - val_loss: 0.3239 - val_accuracy: 0.8687 625/625 [==========] - 47s 75ms/step - loss: 0.3248 - accuracy: 0.8696 Test accuracy: 0.8695999979972839 Time required to fine-tune: 1729.7143523693085
No	No	No	Snowball Stemmer	Epoch 1/5 750/750 [==========] - 266s 351ms/step - loss: 0.4267 - accuracy: 0.7985 - val_loss: 0.3460 - val_accuracy: 0.8493 Epoch 2/5 750/750 [==========] - 257s 343ms/step - loss: 0.1673 - accuracy: 0.9397 - val_loss: 0.3104 - val_accuracy: 0.8695 Epoch 5/5 750/750 [==========] - 258s 344ms/step - loss: 0.1167 - accuracy: 0.9657 - val_loss: 0.3037 - val_accuracy: 0.8790 625/625 [==========] - 45s 73ms/step - loss: 0.2969 - accuracy: 0.8781 Test accuracy: 0.8781499862670898 Time required to fine-tune: 1714.3936071395874

2- Glove

Stop Words	Tokenize	Pos	Dimension	Epoch Results
No	No	No	100	750/750 [==========] - 190s 246ms/step - loss: 0.4559 - accuracy: 0.7809 - val_loss: 0.3434 - val_accuracy: 0.8502 Epoch 2/5 750/750 [==========] - 183s 244ms/step - loss: 0.3137 - accuracy: 0.8662 - val_loss: 0.3127 - val_accuracy: 0.8678 Epoch 3/5 750/750 [==========] - 184s 246ms/step - loss: 0.2607 - accuracy: 0.8920 - val_loss: 0.2990 - val_accuracy: 0.8743 Epoch 4/5 750/750 [==========] - 183s 244ms/step - loss: 0.2167 - accuracy: 0.9153 - val_loss: 0.3503 - val_accuracy: 0.8443 Epoch 5/5 750/750 [==========] - 183s 244ms/step - loss: 0.1760 - accuracy: 0.9340 - val_loss: 0.3137 - val_accuracy: 0.8717 625/625 [==========] - 29s 46ms/step - loss: 0.3171 - accuracy: 0.8644 Test accuracy: 0.8644000291824341
No	No	No	200	Epoch 1/5 750/750 [==========] - 174s 228ms/step - loss: 0.4101 - accuracy: 0.8048 - val_loss: 0.3090 - val_accuracy: 0.8718 Epoch 2/5 750/750 [==========] - 173s 230ms/step - loss: 0.2788 - accuracy: 0.8844 - val_loss: 0.2806 - val_accuracy: 0.8803 Epoch 3/5 750/750 [==========] - 169s 226ms/step - loss: 0.2140 - accuracy: 0.9150 - val_loss: 0.2992 - val_accuracy: 0.8708 Epoch 4/5 750/750 [==========] - 169s 226ms/step - loss: 0.1577 - accuracy: 0.9435 - val_loss: 0.2790 - val_accuracy: 0.8828 Epoch 5/5 750/750 [==========] - 171s 227ms/step - loss: 0.1084 - accuracy: 0.9673 - val_loss: 0.2872 - val_accuracy: 0.8817 625/625 [==========] - 27s 43ms/step - loss: 0.2910 - accuracy: 0.8835 Test accuracy: 0.8835499882698059 Time required to fine-tune: 921.8578188419342
No	No	No	300	Epoch 1/5 750/750 [==========] - 179s 235ms/step - loss: 0.3902 - accuracy: 0.8217 - val_loss: 0.3019 - val_accuracy: 0.8753 Epoch 2/5 750/750 [==========] - 173s 231ms/step - loss: 0.2528 - accuracy: 0.8959 - val_loss: 0.2744 - val_accuracy: 0.8883 Epoch 3/5 750/750 [==========] - 174s 232ms/step - loss: 0.1846 - accuracy: 0.9302 - val_loss: 0.2635 - val_accuracy: 0.8925 Epoch 4/5 750/750 [==========] - 173s 230ms/step - loss: 0.1206 - accuracy: 0.9601 - val_loss: 0.2942 - val_accuracy: 0.8870 Epoch 5/5 750/750 [==========] - 173s 231ms/step - loss: 0.0702 - accuracy: 0.9825 - val_loss: 0.3101 - val_accuracy: 0.8828 625/625 [==========] - 27s 44ms/step - loss: 0.3178 - accuracy: 0.8789 Test accuracy: 0.8789499998092651 Time required to fine-tune: 1006.1718916893005
No	No	No	50	Epoch 1/5 750/750 [==========] - 183s 240ms/step - loss: 0.5043 - accuracy: 0.7503 - val_loss: 0.4002 - val_accuracy: 0.8260 Epoch 2/5 750/750 [==========] - 180s 240ms/step - loss: 0.3671 - accuracy: 0.8388 - val_loss: 0.3576 - val_accuracy: 0.8438 Epoch 3/5 750/750 [==========] - 175s 233ms/step - loss: 0.3198 - accuracy: 0.8652 - val_loss: 0.3614 - val_accuracy: 0.8397 Epoch 4/5 750/750 [==========] - 175s 233ms/step - loss: 0.2858 - accuracy: 0.8797 - val_loss: 0.3447 - val_accuracy: 0.8517 Epoch 5/5 750/750 [==========] - 178s 237ms/step - loss: 0.2565 - accuracy: 0.8972 - val_loss: 0.3349 - val_accuracy: 0.8507 625/625 [==========] - 27s 43ms/step - loss: 0.3350 - accuracy: 0.8525 Test accuracy: 0.8524500131607056 Time required to fine-tune: 955.3397836685181

Hybrid Approaches

1- Afinn + ML

Stop Words	Tokenize	Pos	Stemmer/ Lemmatizer	Vectorizer	Classification Report
No	No	No	No	Tf- Idf Vectorizer	Logistic Regression Accuracy: 0.89005 Logistic Regression Classification Report: precision recall f1-score support negative 0.90 0.88 0.89 9935 positive 0.88 0.90 0.89 10065 accuracy 0.89 20000 macro avg 0.89 0.89 0.89 20000 weighted avg 0.89 0.89 0.89 20000

<table>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Tf- Idf Vectorizer</td>
<td>
Decision Tree Accuracy: 0.71155
Decision Tree Classification Report:
<table>
<tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>negative</td><td>0.71</td><td>0.71</td><td>0.71</td><td>9935</td></tr>
<tr><td>positive</td><td>0.72</td><td>0.71</td><td>0.71</td><td>10065</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.71</td><td>20000</td></tr>
<tr><td>macro avg</td><td>0.71</td><td>0.71</td><td>0.71</td><td>20000</td></tr>
<tr><td>weighted avg</td><td>0.71</td><td>0.71</td><td>0.71</td><td>20000</td></tr>
</table>
Random Forest Accuracy: 0.84345
Random Forest Classification Report:
<table>
<tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>negative</td><td>0.83</td><td>0.86</td><td>0.84</td><td>9935</td></tr>
<tr><td>positive</td><td>0.85</td><td>0.83</td><td>0.84</td><td>10065</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.84</td><td>20000</td></tr>
<tr><td>macro avg</td><td>0.84</td><td>0.84</td><td>0.84</td><td>20000</td></tr>
<tr><td>weighted avg</td><td>0.84</td><td>0.84</td><td>0.84</td><td>20000</td></tr>
</table>
Gradient Boosting Accuracy: 0.81565
Gradient Boosting Classification Report:
<table>
<tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>negative</td><td>0.84</td><td>0.77</td><td>0.81</td><td>9935</td></tr>
<tr><td>positive</td><td>0.79</td><td>0.86</td><td>0.82</td><td>10065</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.82</td><td>20000</td></tr>
<tr><td>macro avg</td><td>0.82</td><td>0.82</td><td>0.82</td><td>20000</td></tr>
<tr><td>weighted avg</td><td>0.82</td><td>0.82</td><td>0.82</td><td>20000</td></tr>
</table>
Neural Network Accuracy: 0.8643
Neural Network Classification Report:
<table>
<tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>negative</td><td>0.87</td><td>0.86</td><td>0.86</td><td>9935</td></tr>
<tr><td>positive</td><td>0.86</td><td>0.87</td><td>0.87</td><td>10065</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.86</td><td>20000</td></tr>
<tr><td>macro avg</td><td>0.86</td><td>0.86</td><td>0.86</td><td>20000</td></tr>
<tr><td>weighted avg</td><td>0.86</td><td>0.86</td><td>0.86</td><td>20000</td></tr>
</table>
KNN Accuracy: 0.7169
KNN Classification Report:
<table>
<tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>negative</td><td>0.73</td><td>0.68</td><td>0.70</td><td>9935</td></tr>
<tr><td>positive</td><td>0.70</td><td>0.75</td><td>0.73</td><td>10065</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.72</td><td>20000</td></tr>
<tr><td>macro avg</td><td>0.72</td><td>0.72</td><td>0.72</td><td>20000</td></tr>
<tr><td>weighted avg</td><td>0.72</td><td>0.72</td><td>0.72</td><td>20000</td></tr>
</table>
</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Count Vectorizer</td>
<td>
Logistic Regression Accuracy: 0.8696
Logistic Regression Classification Report:
<table>
<tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>negative</td><td>0.87</td><td>0.87</td><td>0.87</td><td>9935</td></tr>
<tr><td>positive</td><td>0.87</td><td>0.87</td><td>0.87</td><td>10065</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.87</td><td>20000</td></tr>
<tr><td>macro avg</td><td>0.87</td><td>0.87</td><td>0.87</td><td>20000</td></tr>
<tr><td>weighted avg</td><td>0.87</td><td>0.87</td><td>0.87</td><td>20000</td></tr>
</table>
Decision Tree Accuracy: 0.72205
Decision Tree Classification Report:
<table>
<tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>negative</td><td>0.72</td><td>0.73</td><td>0.72</td><td>9935</td></tr>
<tr><td>positive</td><td>0.73</td><td>0.71</td><td>0.72</td><td>10065</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.72</td><td>20000</td></tr>
<tr><td>macro avg</td><td>0.72</td><td>0.72</td><td>0.72</td><td>20000</td></tr>
<tr><td>weighted avg</td><td>0.72</td><td>0.72</td><td>0.72</td><td>20000</td></tr>
</table>
</td>
</tr>
</table>

					Random Forest Accuracy: 0.8479 Random Forest Classification Report: precision recall f1-score support negative 0.84 0.86 0.85 9935 positive 0.86 0.84 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000 Gradient Boosting Accuracy: 0.81235 Gradient Boosting Classification Report: precision recall f1-score support negative 0.84 0.76 0.80 9935 positive 0.79 0.86 0.82 10065 accuracy 0.81 20000 macro avg 0.82 0.81 0.81 20000 weighted avg 0.81 0.81 0.81 20000 Neural Network Accuracy: 0.8696 Neural Network Classification Report: precision recall f1-score support negative 0.87 0.86 0.87 9935 positive 0.87 0.87 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 KNN Accuracy: 0.6178 KNN Classification Report: precision recall f1-score support negative 0.60 0.69 0.64 9935 positive 0.64 0.54 0.59 10065 accuracy 0.62 20000 macro avg 0.62 0.62 0.62 20000 weighted avg 0.62 0.62 0.62 20000

Yes	Yes	Yes	No	Tf – Idf Vectorizer	Logistic Regression Accuracy: 0.86515 Logistic Regression Classification Report: precision recall f1-score support negative 0.87 0.86 0.86 9935 positive 0.86 0.87 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 Decision Tree Accuracy: 0.7174 Decision Tree Classification Report: precision recall f1-score support negative 0.71 0.72 0.72 9935 positive 0.72 0.71 0.72 10065 accuracy 0.72 20000 macro avg 0.72 0.72 0.72 20000 weighted avg 0.72 0.72 0.72 20000 Random Forest Accuracy: 0.84245 Random Forest Classification Report: precision recall f1-score support negative 0.83 0.85 0.84 9935 positive 0.85 0.83 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 Gradient Boosting Accuracy: 0.805 Gradient Boosting Classification Report: precision recall f1-score support negative 0.84 0.75 0.79 9935 positive 0.78 0.86 0.82 10065 accuracy 0.81 20000 macro avg 0.81 0.80 0.80 20000 weighted avg 0.81 0.81 0.80 20000 Neural Network Accuracy: 0.86515 Neural Network Classification Report: precision recall f1-score support negative 0.87 0.86 0.86 9935 positive 0.86 0.87 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 KNN Accuracy: 0.60745 KNN Classification Report: precision recall f1-score support negative 0.59 0.71 0.64 9935 positive 0.64 0.51 0.57 10065 accuracy 0.61 20000 macro avg 0.61 0.61 0.60 20000 weighted avg 0.61 0.61 0.60 20000

No	No	No	No	Tf-Idf Vectorizer	Improved ML Parameters LogisticRegression(max_iter=1000, C=1.0) Logistic Regression Accuracy: 0.89705 Logistic Regression Classification Report: precision recall f1-score support negative 0.90 0.89 0.90 9935 positive 0.89 0.91 0.90 10065 accuracy 0.90 20000 macro avg 0.90 0.90 0.90 20000 weighted avg 0.90 0.90 0.90 20000 DecisionTreeClassifier(max_depth=10,min_samples_split=5) Decision Tree Accuracy: 0.73175 Decision Tree Classification Report: precision recall f1-score support negative 0.80 0.62 0.70 9935 positive 0.69 0.84 0.76 10065 accuracy 0.73 20000 macro avg 0.74 0.73 0.73 20000 weighted avg 0.74 0.73 0.73 20000 RandomForestClassifier(n_estimators=100, max_depth=10, min_samples_split=5) Random Forest Accuracy: 0.8258 Random Forest Classification Report: precision recall f1-score support negative 0.85 0.79 0.82 9935 positive 0.80 0.86 0.83 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3), Gradient Boosting Accuracy: 0.81655 Gradient Boosting Classification Report: precision recall f1-score support negative 0.84 0.77 0.81 9935 positive 0.79 0.86 0.82 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000 MLPClassifier(max_iter=1000, hidden_layer_sizes=(100,), alpha=0.0001), Neural Network Accuracy: 0.8657 Neural Network Classification Report: precision recall f1-score support negative 0.87 0.86 0.86 9935 positive 0.87 0.87 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000

No	No	No	No	Count Vectorizer	Improved ML Parameters LogisticRegression(max_iter=1000, C=1.0) Logistic Regression Accuracy: 0.8762 Logistic Regression Classification Report: precision recall f1-score support negative 0.88 0.87 0.88 9935 positive 0.88 0.88 0.88 10065 accuracy 0.88 20000 macro avg 0.88 0.88 0.88 20000 weighted avg 0.88 0.88 0.88 20000 DecisionTreeClassifier(max_depth=10,min_samples_split=5) Decision Tree Accuracy: 0.73015 Decision Tree Classification Report: precision recall f1-score support negative 0.78 0.63 0.70 9935 positive 0.69 0.83 0.76 10065 accuracy 0.73 20000 macro avg 0.74 0.73 0.73 20000 weighted avg 0.74 0.73 0.73 20000 RandomForestClassifier(n_estimators=100, max_depth=10, min_samples_split=5) Random Forest Accuracy: 0.82395 Random Forest Classification Report: precision recall f1-score support negative 0.86 0.77 0.81 9935 positive 0.79 0.88 0.83 10065 accuracy 0.82 20000 macro avg 0.83 0.82 0.82 20000 weighted avg 0.83 0.82 0.82 20000 GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3) Gradient Boosting Accuracy: 0.81795 Gradient Boosting Classification Report: precision recall f1-score support negative 0.85 0.77 0.81 9935 positive 0.79 0.86 0.83 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000 MLPClassifier(max_iter=1000, hidden_layer_sizes=(100,), alpha=0.0001) Neural Network Accuracy: 0.892 Neural Network Classification Report: precision recall f1-score support negative 0.90 0.88 0.89 9935 positive 0.89 0.90 0.89 10065 accuracy 0.89 20000 macro avg 0.89 0.89 0.89 20000 weighted avg 0.89 0.89 0.89 20000 MultinomialNB() Naive Bayes Accuracy: 0.85135 Naive Bayes Classification Report: precision recall f1-score support negative 0.85 0.85 0.85 9935 positive 0.85 0.85 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000

<table>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Porter Stemmer</td>
<td>Tf-Idf Vectorizer</td>
<td>Improved ML Parameters
LogisticRegression(max_iter=1000, C=1.0)
Logistic Regression Accuracy: 0.8947
Logistic Regression Classification Report:
precision recall f1-score support

negative 0.90 0.89 0.89 9935
positive 0.89 0.90 0.90 10065

accuracy 0.89 20000
macro avg 0.89 0.89 0.89 20000
weighted avg 0.89 0.89 0.89 20000
DecisionTreeClassifier(max_depth=10,min_samples_split=5)
Decision Tree Accuracy: 0.7223
Decision Tree Classification Report:
precision recall f1-score support

negative 0.80 0.59 0.68 9935
positive 0.68 0.85 0.76 10065

accuracy 0.72 20000
macro avg 0.74 0.72 0.72 20000
weighted avg 0.74 0.72 0.72 20000
RandomForestClassifier(n_estimators=100, max_depth=10, min_samples_split=5)
Random Forest Accuracy: 0.8253
Random Forest Classification Report:
precision recall f1-score support

negative 0.85 0.78 0.82 9935
positive 0.80 0.87 0.83 10065

accuracy 0.83 20000
macro avg 0.83 0.83 0.82 20000
weighted avg 0.83 0.83 0.82 20000
GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3)
Gradient Boosting Accuracy: 0.81215
Gradient Boosting Classification Report:
precision recall f1-score support

negative 0.84 0.77 0.80 9935
positive 0.79 0.85 0.82 10065

accuracy 0.81 20000
macro avg 0.81 0.81 0.81 20000
weighted avg 0.81 0.81 0.81 20000
MLPClassifier(max_iter=1000, hidden_layer_sizes=(100,), alpha=0.0001)
Neural Network Accuracy: 0.8685
Neural Network Classification Report:
precision recall f1-score support

negative 0.87 0.86 0.87 9935
positive 0.87 0.87 0.87 10065

accuracy 0.87 20000
macro avg 0.87 0.87 0.87 20000
weighted avg 0.87 0.87 0.87 20000</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Lancaster Stemmer</td>
<td>Tf-Idf Vectorizer</td>
<td>Improved ML Parameters
LogisticRegression(max_iter=1000, C=1.0)
Logistic Regression Accuracy: 0.8932
Logistic Regression Classification Report:
precision recall f1-score support

negative 0.90 0.88 0.89 9935
positive 0.89 0.90 0.89 10065

accuracy 0.89 20000
macro avg 0.89 0.89 0.89 20000
weighted avg 0.89 0.89 0.89 20000</td>
</tr>
</table>

					DecisionTreeClassifier(max_depth=10,min_samples_split=5) Decision Tree Accuracy: 0.7242 Decision Tree Classification Report: precision recall f1-score support negative 0.80 0.59 0.68 9935 positive 0.68 0.85 0.76 10065 accuracy 0.72 20000 macro avg 0.74 0.72 0.72 20000 weighted avg 0.74 0.72 0.72 20000 RandomForestClassifier(n_estimators=100, max_depth=10, min_samples_split=5) Random Forest Accuracy: 0.8192 Random Forest Classification Report: precision recall f1-score support negative 0.85 0.77 0.81 9935 positive 0.79 0.86 0.83 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000 GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3),\ Gradient Boosting Accuracy: 0.81015 Gradient Boosting Classification Report: precision recall f1-score support negative 0.84 0.77 0.80 9935 positive 0.79 0.85 0.82 10065 accuracy 0.81 20000 macro avg 0.81 0.81 0.81 20000 weighted avg 0.81 0.81 0.81 20000 MLPClassifier(max_iter=1000, hidden_layer_sizes=(100,), alpha=0.0001) Neural Network Accuracy: 0.85525 Neural Network Classification Report: precision recall f1-score support negative 0.85 0.86 0.86 9935 positive 0.86 0.85 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000
No	No	No	Spacy	Tf-Idf Vectorizer	Improved ML Parameters LogisticRegression(max_iter=1000, C=1.0) Logistic Regression Accuracy: 0.89655 Logistic Regression Classification Report: precision recall f1-score support negative 0.90 0.89 0.90 9935 positive 0.89 0.91 0.90 10065 accuracy 0.90 20000 macro avg 0.90 0.90 0.90 20000 weighted avg 0.90 0.90 0.90 20000 DecisionTreeClassifier(max depth=10,min samples split=5) Decision Tree Accuracy: 0.73585 Decision Tree Classification Report: precision recall f1-score support negative 0.76 0.68 0.72 9935 positive 0.71 0.79 0.75 10065 accuracy 0.74 20000 macro avg 0.74 0.74 0.73 20000 weighted avg 0.74 0.74 0.73 20000

					RandomForestClassifier(n_estimators=100, max_depth=10, min_samples_split=5) Random Forest Accuracy: 0.8263 Random Forest Classification Report: precision recall f1-score support negative 0.85 0.78 0.82 9935 positive 0.80 0.87 0.83 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3) Gradient Boosting Accuracy: 0.818 Gradient Boosting Classification Report: precision recall f1-score support negative 0.84 0.78 0.81 9935 positive 0.80 0.86 0.83 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000 MLPClassifier(max_iter=1000, hidden_layer_sizes=(100,), alpha=0.0001) Neural Network Accuracy: 0.86365 Neural Network Classification Report: precision recall f1-score support negative 0.85 0.87 0.86 9935 positive 0.87 0.85 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000

2- Text Blob + ML

Stop Words	Tokenize	Pos	Stemmer/ Lemmatizer	Vectorizer	Classification Report
No	No	No	No	Count Vectorizer	Improved ML Parameters LogisticRegression(max_iter=1000, C=1.0) Logistic Regression Accuracy: 0.8762 Logistic Regression Classification Report: precision recall f1-score support negative 0.88 0.87 0.88 9935 positive 0.88 0.88 0.88 10065 accuracy 0.88 20000 macro avg 0.88 0.88 0.88 20000 weighted avg 0.88 0.88 0.88 20000 DecisionTreeClassifier(max_depth=10,min_samples_split=5) Decision Tree Accuracy: 0.7301 Decision Tree Classification Report: precision recall f1-score support negative 0.79 0.63 0.70 9935 positive 0.69 0.83 0.76 10065 accuracy 0.73 20000 macro avg 0.74 0.73 0.73 20000 weighted avg 0.74 0.73 0.73 20000

					RandomForestClassifier(n_estimators=100, max_depth=10, min_samples_split=5) Random Forest Accuracy: 0.82745 Random Forest Classification Report: precision recall f1-score support negative 0.85 0.79 0.82 9935 positive 0.81 0.87 0.83 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3) Gradient Boosting Accuracy: 0.81795 Gradient Boosting Classification Report: precision recall f1-score support negative 0.85 0.77 0.81 9935 positive 0.79 0.86 0.83 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000 MLPClassifier(max_iter=1000, hidden_layer_sizes=(100,), alpha=0.0001) Neural Network Accuracy: 0.8901 Neural Network Classification Report: precision recall f1-score support negative 0.89 0.88 0.89 9935 positive 0.89 0.90 0.89 10065 accuracy 0.89 20000 macro avg 0.89 0.89 0.89 20000 weighted avg 0.89 0.89 0.89 20000 MultinomialNB() Naive Bayes Accuracy: 0.85135 Naive Bayes Classification Report: precision recall f1-score support negative 0.85 0.85 0.85 9935 positive 0.85 0.85 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000
No	No	No	No	Tf – Idf Vectorizer	Improved ML Parameters LogisticRegression(max_iter=1000, C=1.0) Logistic Regression Accuracy: 0.89705 Logistic Regression Classification Report: precision recall f1-score support negative 0.90 0.89 0.90 9935 positive 0.89 0.91 0.90 10065 accuracy 0.90 20000 macro avg 0.90 0.90 0.90 20000 weighted avg 0.90 0.90 0.90 20000 DecisionTreeClassifier(max_depth=10,min_samples_split=5) Decision Tree Accuracy: 0.7317 Decision Tree Classification Report: precision recall f1-score support negative 0.80 0.62 0.70 9935 positive 0.69 0.84 0.76 10065 accuracy 0.73 20000 macro avg 0.74 0.73 0.73 20000 weighted avg 0.74 0.73 0.73 20000

					RandomForestClassifier(n_estimators=100, max_depth=10, min_samples_split=5) Random Forest Accuracy: 0.82445 Random Forest Classification Report: precision recall f1-score support negative 0.86 0.78 0.81 9935 positive 0.80 0.87 0.83 10065 accuracy 0.82 20000 macro avg 0.83 0.82 0.82 20000 weighted avg 0.83 0.82 0.82 20000 GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3) Gradient Boosting Accuracy: 0.8166 Gradient Boosting Classification Report: precision recall f1-score support negative 0.84 0.77 0.81 9935 positive 0.79 0.86 0.82 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000 MLPClassifier(max_iter=1000, hidden_layer_sizes=(100,), alpha=0.0001) Neural Network Accuracy: 0.86825 Neural Network Classification Report: precision recall f1-score support negative 0.87 0.86 0.87 9935 positive 0.87 0.87 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 MultinomialNB() Naive Bayes Accuracy: 0.86895 Naive Bayes Classification Report: precision recall f1-score support negative 0.87 0.86 0.87 9935 positive 0.86 0.88 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000

3- Vader + ML

Stop Words	Tokenize	Pos	Stemmer/ Lemmatizer	Vectorizer	Classification Report
No	No	No	No	Count Vectorizer	Improved ML Parameters LogisticRegression(max_iter=1000, C=1.0) Logistic Regression Accuracy: 0.8762 Logistic Regression Classification Report: precision recall f1-score support negative 0.88 0.87 0.88 9935 positive 0.88 0.88 0.88 10065 accuracy 0.88 20000 macro avg 0.88 0.88 0.88 20000 weighted avg 0.88 0.88 0.88 20000 DecisionTreeClassifier(max_depth=10,min_samples_split=5)

					Decision Tree Accuracy: 0.72985 Decision Tree Classification Report: precision recall f1-score support negative 0.79 0.63 0.70 9935 positive 0.69 0.83 0.76 10065 accuracy 0.73 20000 macro avg 0.74 0.73 0.73 20000 weighted avg 0.74 0.73 0.73 20000 RandomForestClassifier(n_estimators=100, max_depth=10, min_samples_split=5) Random Forest Accuracy: 0.8256 Random Forest Classification Report: precision recall f1-score support negative 0.86 0.78 0.82 9935 positive 0.80 0.87 0.83 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3) Gradient Boosting Accuracy: 0.81795 Gradient Boosting Classification Report: precision recall f1-score support negative 0.85 0.77 0.81 9935 positive 0.79 0.86 0.83 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000 MLPClassifier(max_iter=1000, hidden_layer_sizes=(100,), alpha=0.0001) Neural Network Accuracy: 0.8874 Neural Network Classification Report: precision recall f1-score support negative 0.89 0.89 0.89 9935 positive 0.89 0.89 0.89 10065 accuracy 0.89 20000 macro avg 0.89 0.89 0.89 20000 weighted avg 0.89 0.89 0.89 20000 MultinomialNB() Naive Bayes Accuracy: 0.85135 Naive Bayes Classification Report: precision recall f1-score support negative 0.85 0.85 0.85 9935 positive 0.85 0.85 0.85 10065 accuracy 0.85 20000 macro avg 0.85 0.85 0.85 20000 weighted avg 0.85 0.85 0.85 20000
No	No	No	No	Tf – Idf Vectorizer	Improved ML Parameters LogisticRegression(max_iter=1000, C=1.0) Logistic Regression Accuracy: 0.89705 Logistic Regression Classification Report: precision recall f1-score support negative 0.90 0.89 0.90 9935 positive 0.89 0.91 0.90 10065 accuracy 0.90 20000 macro avg 0.90 0.90 0.90 20000 weighted avg 0.90 0.90 0.90 20000

					DecisionTreeClassifier(max_depth=10,min_samples_split=5) Decision Tree Accuracy: 0.73165 Decision Tree Classification Report: precision recall f1-score support negative 0.80 0.62 0.70 9935 positive 0.69 0.84 0.76 10065 accuracy 0.73 20000 macro avg 0.74 0.73 0.73 20000 weighted avg 0.74 0.73 0.73 20000 RandomForestClassifier(n_estimators=100, max_depth=10, min_samples_split=5) Random Forest Accuracy: 0.82395 Random Forest Classification Report: precision recall f1-score support negative 0.85 0.78 0.82 9935 positive 0.80 0.86 0.83 10065 accuracy 0.82 20000 macro avg 0.83 0.82 0.82 20000 weighted avg 0.83 0.82 0.82 20000 GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3) Gradient Boosting Accuracy: 0.8166 Gradient Boosting Classification Report: precision recall f1-score support negative 0.84 0.77 0.81 9935 positive 0.79 0.86 0.82 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000 MLPClassifier(max_iter=1000, hidden_layer_sizes=(100,), alpha=0.0001) Neural Network Accuracy: 0.8596 Neural Network Classification Report: precision recall f1-score support negative 0.88 0.83 0.85 9935 positive 0.84 0.89 0.86 10065 accuracy 0.86 20000 macro avg 0.86 0.86 0.86 20000 weighted avg 0.86 0.86 0.86 20000 MultinomialNB() Naive Bayes Accuracy: 0.86895 Naive Bayes Classification Report: precision recall f1-score support negative 0.87 0.86 0.87 9935 positive 0.86 0.88 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000
No	Yes	No	No	Tf – Idf Vectorizer	Improved ML Parameters LogisticRegression(max_iter=1000, C=1.0) Logistic Regression Accuracy: 0.897 Logistic Regression Classification Report: precision recall f1-score support negative 0.90 0.89 0.90 9935 positive 0.89 0.90 0.90 10065 accuracy 0.90 20000 macro avg 0.90 0.90 0.90 20000 weighted avg 0.90 0.90 0.90 20000

					DecisionTreeClassifier(max_depth=10,min_samples_split=5) Decision Tree Accuracy: 0.7329 Decision Tree Classification Report: precision recall f1-score support negative 0.80 0.62 0.70 9935 positive 0.69 0.85 0.76 10065 accuracy 0.73 20000 macro avg 0.75 0.73 0.73 20000 weighted avg 0.74 0.73 0.73 20000 RandomForestClassifier(n_estimators=100, max_depth=10, min_samples_split=5) Random Forest Accuracy: 0.82395 Random Forest Classification Report: precision recall f1-score support negative 0.86 0.77 0.81 9935 positive 0.80 0.87 0.83 10065 accuracy 0.82 20000 macro avg 0.83 0.82 0.82 20000 weighted avg 0.83 0.82 0.82 20000 GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3) Gradient Boosting Accuracy: 0.8182 Gradient Boosting Classification Report: precision recall f1-score support negative 0.85 0.77 0.81 9935 positive 0.79 0.86 0.83 10065 accuracy 0.82 20000 macro avg 0.82 0.82 0.82 20000 weighted avg 0.82 0.82 0.82 20000 MLPClassifier(max_iter=1000, hidden_layer_sizes=(100,), alpha=0.0001) Neural Network Accuracy: 0.86785 Neural Network Classification Report: precision recall f1-score support negative 0.87 0.87 0.87 9935 positive 0.87 0.87 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000 MultinomialNB() Naive Bayes Accuracy: 0.8696 Naive Bayes Classification Report: precision recall f1-score support negative 0.88 0.86 0.87 9935 positive 0.86 0.88 0.87 10065 accuracy 0.87 20000 macro avg 0.87 0.87 0.87 20000 weighted avg 0.87 0.87 0.87 20000

4- Afinn + Vader + TextBlob + ML

Stop Words	Tokenize	Pos	Stemmer/ Lemmatizer	Vectorizer	Classification Report
No	No	No	No	Tf – Idf Vectorizer	Improved ML Parameters LogisticRegression(max_iter=1000, C=1.0)

					Logistic Regression Accuracy: 0.8919 Logistic Regression Classification Report: precision recall f1-score support negative 0.89 0.89 0.89 9935 positive 0.89 0.90 0.89 10065 accuracy 0.89 20000 macro avg 0.89 0.89 0.89 20000 weighted avg 0.89 0.89 0.89 20000 DecisionTreeClassifier(max_depth=10,min_samples_split=5) Decision Tree Accuracy: 0.7758 Decision Tree Classification Report: precision recall f1-score support negative 0.79 0.74 0.77 9935 positive 0.76 0.81 0.78 10065 accuracy 0.78 20000 macro avg 0.78 0.78 0.78 20000 weighted avg 0.78 0.78 0.78 20000 RandomForestClassifier(n_estimators=100, max_depth=10, min_samples_split=5) Random Forest Accuracy: 0.83625 Random Forest Classification Report: precision recall f1-score support negative 0.85 0.82 0.83 9935 positive 0.83 0.85 0.84 10065 accuracy 0.84 20000 macro avg 0.84 0.84 0.84 20000 weighted avg 0.84 0.84 0.84 20000 GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3) Gradient Boosting Accuracy: 0.82955 Gradient Boosting Classification Report: precision recall f1-score support negative 0.83 0.83 0.83 9935 positive 0.83 0.83 0.83 10065 accuracy 0.83 20000 macro avg 0.83 0.83 0.83 20000 weighted avg 0.83 0.83 0.83 20000 MLPClassifier(max_iter=1000, hidden_layer_sizes=(100,), alpha=0.0001) Neural Network Accuracy: 0.8751 Neural Network Classification Report: precision recall f1-score support negative 0.87 0.87 0.87 9935 positive 0.88 0.88 0.88 10065 accuracy 0.88 20000 macro avg 0.88 0.88 0.88 20000 weighted avg 0.88 0.88 0.88 20000

5- Vader + Glove100d

Stop Words	Tokenize	Pos	Stemmer/ Lemmatizer	Epoch Results
No	No	No	No	Epoch 1/5 188/188 [==========] - 72s 360ms/step - loss: 0.5750 - accuracy: 0.6756 - val_loss: 0.4333 - val_accuracy: 0.7975 Epoch 2/5 188/188 [==========] - 66s 353ms/step - loss: 0.3555 - accuracy: 0.8497 - val_loss: 0.4162 - val_accuracy: 0.8127 Epoch 3/5 188/188 [==========] - 66s 352ms/step - loss: 0.2940 - accuracy: 0.8798 - val_loss: 0.3496 - val_accuracy: 0.8513 Epoch 4/5 188/188 [==========] - 66s 351ms/step - loss: 0.2606 - accuracy: 0.8960 - val_loss: 0.3808 - val_accuracy: 0.8420 Epoch 5/5 188/188 [==========] - 66s 352ms/step - loss: 0.2338 - accuracy: 0.9108 - val_loss: 0.3473 - val_accuracy: 0.8555 625/625 [==========] - 35s 56ms/step - loss: 0.3349 - accuracy: 0.8558 Accuracy: 85.58%

6- Vader + Glove200d

Stop Words	Tokenize	Pos	Stemmer/ Lemmatizer	Epoch Results
No	No	No	No	Epoch 1/5 188/188 [==========] - 67s 337ms/step - loss: 0.5512 - accuracy: 0.6990 - val_loss: 0.3853 - val_accuracy: 0.8308 Epoch 2/5 188/188 [==========] - 61s 327ms/step - loss: 0.3508 - accuracy: 0.8508 - val_loss: 0.5723 - val_accuracy: 0.7973 Epoch 3/5 188/188 [==========] - 62s 332ms/step - loss: 0.2947 - accuracy: 0.8784 - val_loss: 0.3388 - val_accuracy: 0.8552 Epoch 4/5 188/188 [==========] - 61s 327ms/step - loss: 0.2569 - accuracy: 0.8999 - val_loss: 0.3649 - val_accuracy: 0.8533 Epoch 5/5 188/188 [==========] - 63s 334ms/step - loss: 0.2304 - accuracy: 0.9115 - val_loss: 0.3473 - val_accuracy: 0.8520 625/625 [==========] - 33s 53ms/step - loss: 0.3356 - accuracy: 0.8541 Accuracy: 85.41%

7- Vader + Glove300d

Stop Words	Tokenize	Pos	Stemmer/ Lemmatizer	Epoch Results
No	No	No	No	Epoch 1/5 188/188 [==========] - 67s 338ms/step - loss: 0.5814 - accuracy: 0.6700 - val_loss: 0.3913 - val_accuracy: 0.8300 Epoch 2/5 188/188 [==========] - 62s 332ms/step - loss: 0.3591 - accuracy: 0.8457 - val_loss: 0.3781 - val_accuracy: 0.8305 Epoch 3/5 188/188 [==========] - 62s 332ms/step - loss: 0.2954 - accuracy: 0.8785 - val_loss: 0.4211 - val_accuracy: 0.8042 Epoch 4/5 188/188 [==========] - 63s 333ms/step - loss: 0.2598 - accuracy: 0.8963 - val_loss: 0.3564 - val_accuracy: 0.8538 Epoch 5/5 188/188 [==========] - 63s 333ms/step - loss: 0.2309 - accuracy: 0.9104 - val_loss: 0.3470 - val_accuracy: 0.8532 625/625 [==========] - 33s 52ms/step - loss: 0.3347 - accuracy: 0.8566 Accuracy: 85.66%

8- Afinn + Glove100d

Stop Words	Tokenize	Pos	Stemmer/ Lemmatizer	Epoch Results
No	No	No	No	Epoch 1/5 188/188 [==========] - 72s 360ms/step - loss: 0.5705 - accuracy: 0.6835 - val_loss: 0.3759 - val_accuracy: 0.8360 Epoch 2/5 188/188 [==========] - 68s 361ms/step - loss: 0.3568 - accuracy: 0.8473 - val_loss: 0.5320 - val_accuracy: 0.8112 Epoch 3/5 188/188 [==========] - 67s 357ms/step - loss: 0.2948 - accuracy: 0.8795 - val_loss: 0.3460 - val_accuracy: 0.8498 Epoch 4/5 188/188 [==========] - 67s 357ms/step - loss: 0.2547 - accuracy: 0.9004 - val_loss: 0.3391 - val_accuracy: 0.8550 Epoch 5/5 188/188 [==========] - 67s 354ms/step - loss: 0.2305 - accuracy: 0.9090 - val_loss: 0.4364 - val_accuracy: 0.8397 625/625 [==========] - 34s 55ms/step - loss: 0.4242 - accuracy: 0.8413 Accuracy: 84.13%

Conclusion:

In conclusion, our comprehensive analysis of sentiment analysis models on the IMDb dataset underscores the complexity and variability inherent in sentiment analysis tasks. Through experimentation with a diverse array of machine learning models, pre-trained language models, lexicon-based approaches, and word embeddings, we have demonstrated that there is no one-size-fits-all solution for sentiment analysis. The effectiveness of sentiment analysis models is highly dependent on various factors such as the choice of preprocessing techniques, vectorization methods, and feature extraction strategies. Our findings highlight the importance of systematically exploring and evaluating different combinations of models and preprocessing pipelines to achieve optimal performance. It is imperative to recognize that what works well for one dataset or application scenario may not necessarily generalize to others. Therefore, practitioners should approach sentiment analysis tasks with careful consideration and experimentation, tailoring their approach to the specific characteristics and requirements of the dataset and the problem at hand.

Best Model:

Finetuning Roberta

Accuracy:

```
Accuracy: 0.95255
F1-score: 0.9531149646756584
Confusion matrix:
 [[9485  530]
 [ 419 9646]]
```

Classification Report:

	precision	recall	f1-score	support
negative	0.96	0.95	0.95	9935
positive	0.95	0.96	0.95	10065
accuracy			0.95	20000
macro avg	0.95	0.95	0.95	20000
weighted avg	0.95	0.95	0.95	20000

Time required to fine-tune: 6715.162095785141s

Code:

```
import torch
from torch.utils.data import Dataset, DataLoader
from transformers import RobertaTokenizerFast, RobertaForSequenceClassification#, AdamW
from sklearn.model_selection import train_test_split
import pandas as pd
from torch.optim import AdamW

import time
# Record start time
start_time = time.time()

# Load data
data = pd.read_csv('/kaggle/input/plmdataset/train.csv/train.csv')
data['sentiment'] = data['sentiment'].map({'positive': 1, 'negative': 0})
reviews = data['review'].tolist()
labels = data['sentiment'].tolist()  # assuming sentiment is encoded as 0 (negative) and 1 (positive)

# Split data into training and validation sets
train_texts, val_texts, train_labels, val_labels = train_test_split(reviews, labels, test_size=0.2)

# Initialize tokenizer
tokenizer = RobertaTokenizerFast.from_pretrained('roberta-base')

# Tokenize data
train_encodings = tokenizer(train_texts, truncation=True, padding=True, max_length=512)
val_encodings = tokenizer(val_texts, truncation=True, padding=True, max_length=512)

# Create torch dataset
class ReviewDataset(Dataset):
    def _init_(self, encodings, labels):
```

```
        self.encodings = encodings
        self.labels = labels

    def __getitem__(self, idx):
        item = {key: torch.tensor(val[idx]) for key, val in self.encodings.items()}
        item['labels'] = torch.tensor(self.labels[idx])
        return item

    def __len__(self):
        return len(self.labels)

# Create dataloaders
train_dataset = ReviewDataset(train_encodings, train_labels)
val_dataset = ReviewDataset(val_encodings, val_labels)

train_loader = DataLoader(train_dataset, batch_size=16, shuffle=True)
val_loader = DataLoader(val_dataset, batch_size=16, shuffle=False)

# Initialize model
model = RobertaForSequenceClassification.from_pretrained('roberta-base', num_labels=2)
model = model.to('cuda')  # if GPU is available

# Initialize optimizer
optimizer = AdamW(model.parameters(), lr=1e-5)

# Training loop
for epoch in range(3):  # number of epochs
    model.train()
    for batch in train_loader:
        optimizer.zero_grad()
        input_ids = batch['input_ids'].to('cuda')
        attention_mask = batch['attention_mask'].to('cuda')
        labels = batch['labels'].to('cuda')
        outputs = model(input_ids, attention_mask=attention_mask, labels=labels)
        loss = outputs.loss
        loss.backward()
        optimizer.step()

# Save the model
model.save_pretrained('sentiment_model_RoBERTa')

# Record end time
end_time = time.time()

print("Time required to fine-tune: ", end_time - start_time)

from sklearn.metrics import accuracy_score, f1_score, confusion_matrix
from transformers import RobertaTokenizerFast, RobertaForSequenceClassification
from torch.utils.data import DataLoader, Dataset
import pandas as pd
import torch

# Load the model
model = RobertaForSequenceClassification.from_pretrained('sentiment_model_RoBERTa')
model = model.to('cuda')  # if GPU is available

# Load validation data
val_data = pd.read_csv('/kaggle/input/plmdataset/test.csv/test.csv')
val_texts = val_data['review'].tolist()
val_labels = val_data['sentiment'].map({'positive': 1, 'negative': 0}).tolist()  # convert sentiment to numeric

# Initialize tokenizer
```

```
tokenizer = RobertaTokenizerFast.from_pretrained('roberta-base')

# Tokenize data
val_encodings = tokenizer(val_texts, truncation=True, padding=True, max_length=512)

# Create torch dataset for validation
class ReviewDataset(Dataset):
    def _init_(self, encodings, labels):
        self.encodings = encodings
        self.labels = labels

    def _getitem_(self, idx):
        item = {key: torch.tensor(val[idx]) for key, val in self.encodings.items()}
        item['labels'] = torch.tensor(self.labels[idx])
        return item

    def _len_(self):
        return len(self.labels)


val_dataset = ReviewDataset(val_encodings, val_labels)
val_loader = DataLoader(val_dataset, batch_size=16, shuffle=False)

# Evaluate the model
model.eval()
predictions = []
true_labels = []
for batch in val_loader:
    input_ids = batch['input_ids'].to('cuda')
    attention_mask = batch['attention_mask'].to('cuda')
    labels = batch['labels'].to('cuda')

    with torch.no_grad():
        outputs = model(input_ids, attention_mask=attention_mask)

    logits = outputs.logits
    predicted_labels = torch.argmax(logits, dim=1).cpu().numpy()
    predictions.extend(predicted_labels)
    true_labels.extend(labels.cpu().numpy())

# Calculate metrics
accuracy = accuracy_score(true_labels, predictions)
f1 = f1_score(true_labels, predictions)
conf_matrix = confusion_matrix(true_labels, predictions)

print(f'Accuracy: {accuracy}')
print(f'F1-score: {f1}')
print(f'Confusion matrix:\n {conf_matrix}')
```