Group9Phase4 Project

Sentiment Analysis of Amazon Fine Food Reviews

Navigating the Repository

There are a total of 4 branches (main, team, teamb, teamc), and the final complete project is in the main branch. The team(a,b,c) branches were created for task allocation for pair-wise collaborations for the group members; however due to constraints, we eventually abandoned the branches and worked together on the main branch.

****

****

***

Project Introduction

The Project we are working on is based on a dataset from Amazon's Fine Food Reviews, The goal is to build a model(s) that can accurately determine sentiment from a review in text form and classify it as either positive, negative, or neutral.

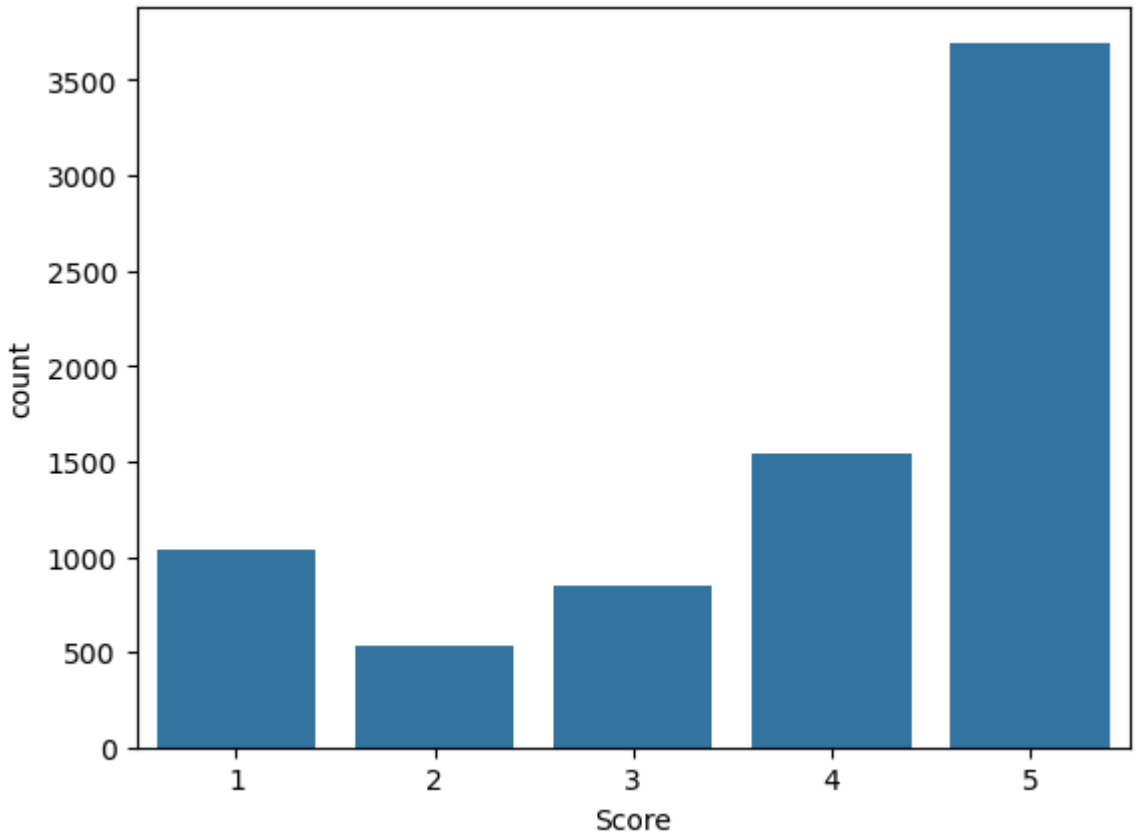
With the model(s) chosen, one can use them to rate reviews quickly and improve them for much better performance or tune them to determine the sentiment on a comment from other sources.

Analysis was conducted on;

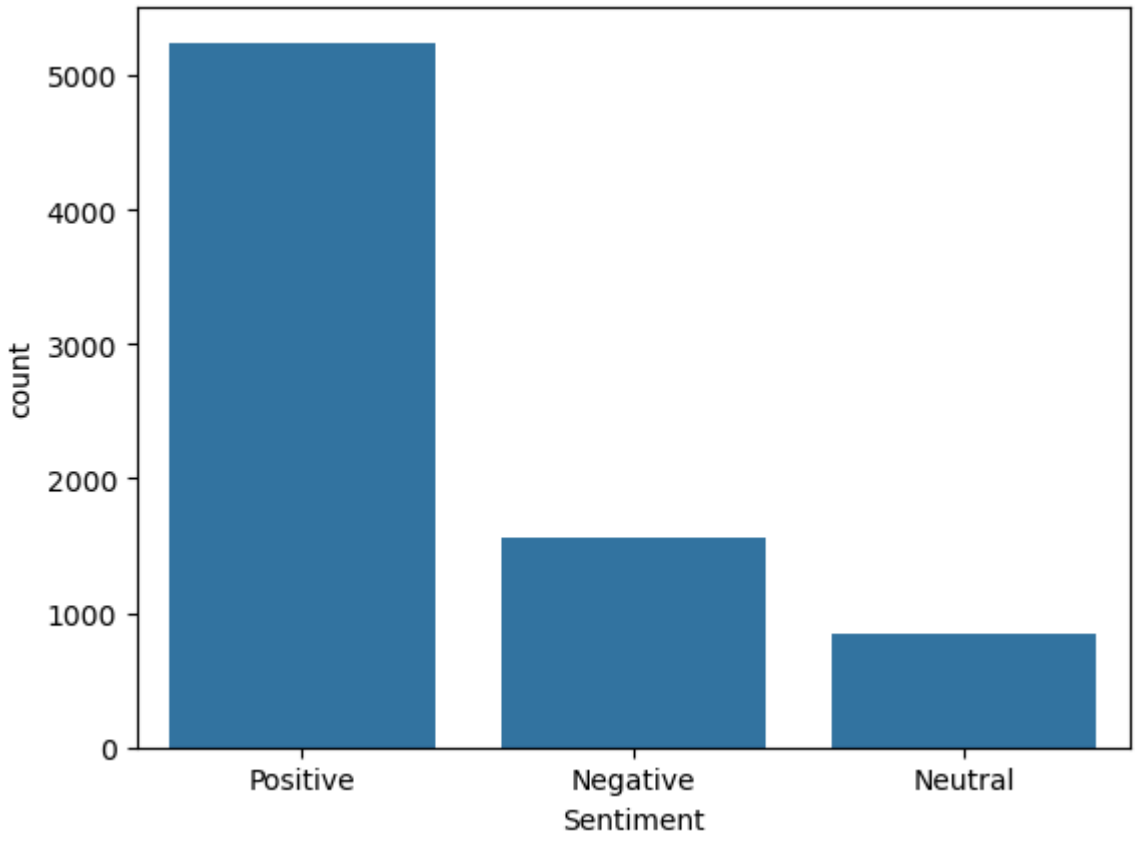
1. Reviews(Text)
2. Ratings(Scores)
3. Sentiments(Positive, Negative, Neutral)

Packages Used pandas, seaborn, matplotlib.pyplot, numpy, warnings, sklearn.linear_model(LogisticRegression), sklearn.metrics(accuracy_score, recall_score, precision_score, f1_score, classification_report, confusion_matrix), sklearn.model_selection(train_test_split), sklearn.model_selection(GridSearchCV), sklearn.ensemble(RandomForestClassifier), sklearn.preprocessing(LabelEncoder), nltk(word_tokenize, stopwords, WordNetLemmatizer, TreebankWordTokenizer), sklearn.feature_extraction.text(TfidfVectorizer,CountVectorizer), sklearn.naive_bayes(MultinomialNB), xgboost(XGBClassifier), tensorflow.keras(models,layers), transformers(DistilBertTokenizer, DistilBertForSequenceClassification, Trainer, TrainingArguments).

Insights

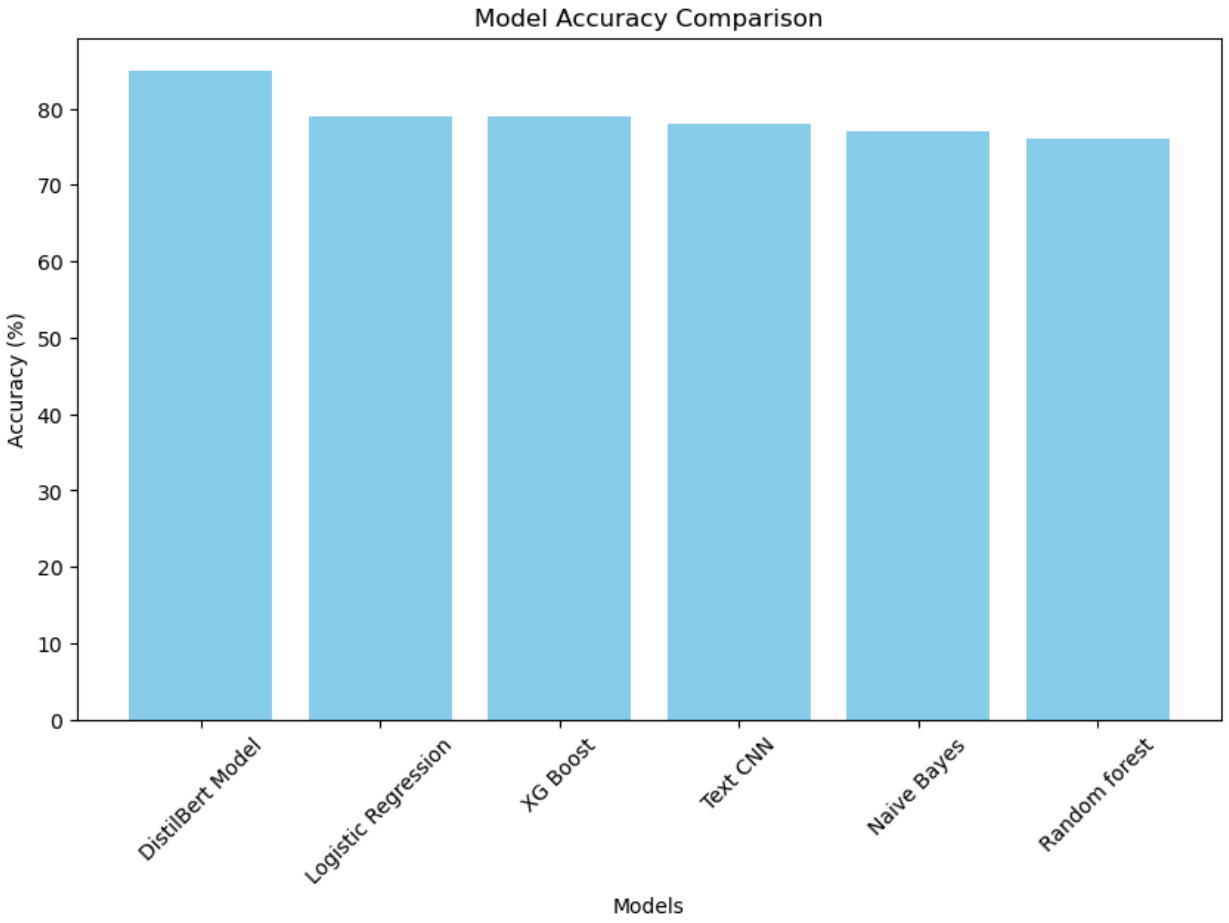


We can see the distribution of Ratings(Scores)



We can see the distribution of sentiments.

Model Performance by overall accuracy.



Conclusion

At face value, the DistillBERT model is the best model; however, it can be further tuned and improved if computational resources are not constrained, however. If they are, then the Logistic regression model is a suitable choice that may suffice in balancing accuracy and computing power.