NLP220 Assignment 1

Steven Au sttau

October 27,2023

1 Part A

The features I checked were testing Bag of Words (BoW) and Term Frequency - Inverse Document Frequency (TF-IDF) with the review corpus or the summary corpus on a Naive Bayes (NB) and Support Vector Machine (SVM) model.

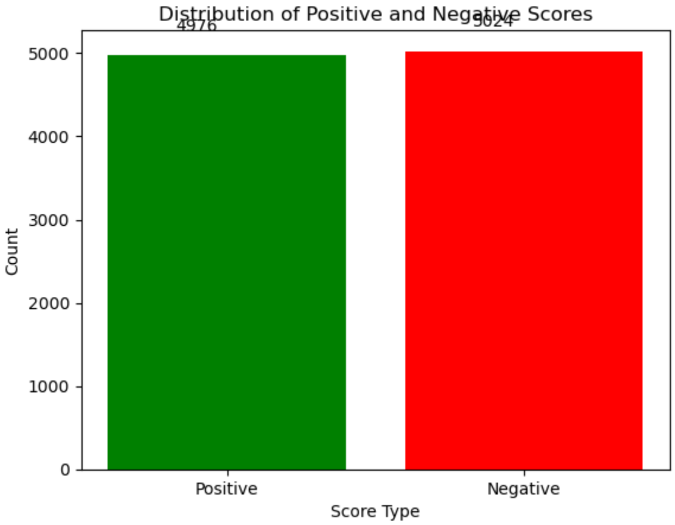

Figure 1: Random sampling for a smaller test set

Summary

Best Model for Accuracy: SVM with `tfidf_review` **Best Model for True Positives:** SVM with `tfidf_review`
Best Model for True Negatives: MNB with `count_review`

In this case, the SVM model trained on the `tfidf_review` feature provides the best balance between accuracy, true positives, and true negatives. SVM was

Table 1: Model Performance Metrics

Model	Accuracy	True Positives (TP)	True Negatives (TN)
SVM (tfidf_review)	0.868	432	436
MNB (tfidf_review)	0.839	404	435
SVM (tfidf_summary)	0.806	396	410
MNB (tfidf_summary)	0.786	376	410
SVM (count_review)	0.812	408	404
MNB (count_review)	0.82	374	446
SVM (count_summary)	0.8	387	413
MNB (count_summary)	0.79	376	414

able to model better than MNB slightly. SVM is better at capturing non-linear data and maximizing hyperplanes giving it a slight edge over al Naive Bayes.

2 Part B

I created a unigram-to-bigram model on a random sample set for a total corpus of 1000 that tokenized location, organization, and person as a single token. I set the max corpus size to 5000 and performed BoW and TF-IDF for each of the 5 models.

I ran a unigram-to-trigram model and it performed worse than just unigrams-to-bigram. BoW performed better in KNN and Logistic regression but, overall TF-IDF did better marginally in the rest of the other models. TF-IDF is better at normalizing the data across documents and thus reducing noise. In most instances, TF-IDF performed better than a simple BoW embedding.

Table 2: Accuracies of Different Models using BoW and TF-IDF

Model	BoW Accuracy (%)	TF-IDF Accuracy (%)
Naive Bayes	79.00	80.00
K-Nearest Neighbors	59.00	53.00
Support Vector Machine	73.00	82.00
Logistic Regression	84.00	83.00
Random Forest	79.00	81.00

Table 3: Best Models using BoW and TF-IDF

Feature Extraction Method	Best Model (Accuracy %)
BoW	Logistic Regression (84.00%)
TF-IDF	Support Vector Machine (82.00%)