The multinomial Naive Bayes model provided by sklearn package is the most suitable for classification with discrete features (e.g., word counts for text classification). As a result, we should use it for our sentiment analysis. It easily out-perform the simple Naïve Bayes model that I built myself.

A picture containing table

Description automatically generated

When applying model to the original Yelp test.csv data, the accuracy is around 90%;

When applying the model to the generated data, since we basically generated using the same intrinsic logic, it has the highest accuracy, 95%;

When applying the model to the real-world data (the fine food review from Amazon), the accuracy is around 70%.