Sentiment Analysis Project Writeup

By: Divya Uppal

May 2022

Abstract:

The goal of this project is to predict the rating of an Amazon product by analyzing the sentiment of the review given by the customer. The model will be used to predict reviews provided by customers on different portals which inturn will be input data to manage the brand and reputation of the product. Both rule based and machine learning sentiment analysis models were evaluated on the data. For Rule based sentiment analyzer Vader and TextBlob were evaluated whereas for Machine Learning based analyser Logistic regression, Naive Bayes and Random Forest were explored.

Impact:

Identifying both happy and unhappy customers is very helpful for the business. We can market several products of interest to happy customers, whereas for unsatisfied customers, various steps can be taken to identify their grievance and processes be put in place to improve product, relationship with customers etc so that we do not lose the customer to the competition.

Design:

Data:

Data was collected from *https://www.kaggle.com/datasets/bittlingmayer/amazonreviews*.
Training data has 3399508 records.. For the scope of the project, the model was trained on .003% of training data with customer review length ranging from 56 words to 1007 words. Dataset is balanced with an equal proportion of positive and negative reviews.

Algorithm:

Data was pre-processed to remove all punctuations, words smaller than 3 letters, stopwords etc. All reviews were converted to lowercase and only alphanumeric characters in the review were analyzed. Performance of data was evaluated using both Count vectorizer and TF-IDF tokenization methods. Performance of lemmatization was also evaluated on the data

Models:

Rule Based Sentiment Analyzer

1. Vader
2. TextBlob

Machine Learning Models

1. Logistic Regression
2. Random Forest

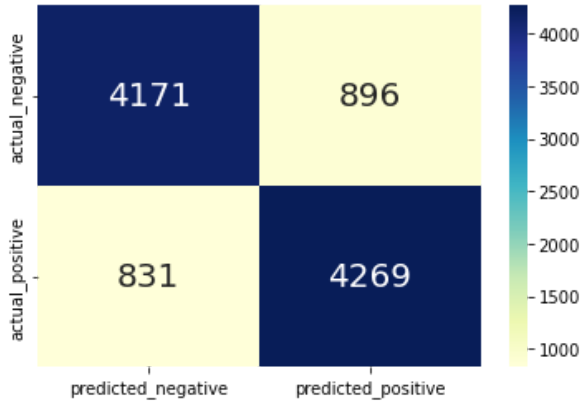
3. Naive Bayes

Both Rule based and Machine Learning models were evaluated before settling in for Logistic Regression Model with an accuracy of .830. Lemmatization was not used on the data and TF-IDF was used to tokenize the data.

Comparison of Various Models:

	Vader	Text Blob	Logistic Regression	Logistic Regression & Lemmatization	Random Forest	Naive Bayes
Accuracy	.688	.662	.830	.817	.821	.806
Precision	.628	.604	.827	.816	.817	.814
Recall	.928	.945	.837	.820	.830	.796
F1 Score	.749	.737	.832	.818	.823	.805

Confusion Matrix:



Tools:

- Scikit-learn for modeling
- Numpy and Pandas for data manipulation
- Nltk and spaCy for text analysis
- Vader and TextBlob for Sentiment Analysis
- Wordcloud for Visual Representation