Amazon Sentiment Analysis

A sentiment prediction model

By Suraj Kumar Prajapati

Introduction

What is sentiment analysis ?

Sentiment analysis is a field of study that aims to computationally identify and classify subjective information. This technique has significant practical applications in areas such as market research, social media monitoring, and customer service, among others.

How sentiment analysis is useful ?

By using natural language processing and machine learning algorithms, sentiment analysis allows us to gain valuable insights into people's opinions, emotions, and attitudes toward specific topics, products, or services.

Prediction of sentiment of the collected data of product from the amazon online store is the task of the this model.

Example

"I really like the new design of your website!" → Positive
"I'm not sure if I like the new design" → Neutral
"The new design is awful!" → Negative

Business Aspect

There are more than 3.5 billion active social media users; that's 45% of the world's population. Every minute users send over 500,000 Tweets and post 510,000 Facebook comments, and a large amount of these messages contain valuable business insights about how customers feel towards products, brands and services.

Boost Business Performance and Strategy

Take Quick Action Against Poor Customer Experiences

Data Driven Support Conversation

Resolve Customer Queries in Real Time

About Data

For the **sentiment analysis model**, I used dataset of reviews of **musical instrument** having **10261** observations and 9 features from amazon.

Data contains **1429** unique customers reviews and ratings Dataset have reviews, which is a short statement about the product and rating provided to the product by that customer.

There are two columns of time, review time and unix review time. It also contains the name of the reviewer and the asin number of the product, which is the unique id for a product. There are reviews of total **900** unique products.

Summary of the review is also available in the form of new column.

Follow the link here for data set.

Pathway

01

Review Cleaning using NLTK

Cleaning of reviews to remove the unimportant words

02

Preprocessing

Preparation of data to fit in the models

03

Model Generation

Creation of several models

04

Evaluation

Evaluate the model and do important chances to improve accuracy

Review Cleaning

Sentence Irregularities

01

Links
Links present in the reviews are not useful for analysis

02

Stopwords
There are lot of repetitive words which are not useful (e.g. Helping Verb)

03

HTML Tags
During web-scrapping html tags may also present which are not useful

04

Alpha-Numeric
Numbers and special character are not useful

05

Punctuation
Punctuation in the sentences are useless

06

New Line Character
Change of line is not matter for analysis

Preprocessing

01

Lemmatization/Stemming

It helps to bring the words into the original form. (e.g. Healthy to Health)

02

Vectorization

Used Tf-idf Vectorizer to convert words into a vector, which the input for the model.

03

Encoding

It used to convert categorical variable into numerical variables

04

Smote Oversampling

It is used to generate artificial samples to remove bias.

Vectorization

" The Quick Brown Fox Jump Over A Lazy Dog "

Tf-Idf Method

The Vectorization can be done in two ways, on the basis of picking of words at a time

Mono-Gram

In this method, individual word of the sentence is vectorize into vector. For this purpose, the n-gram parameter is set to be 1.

Bi-Gram

The method uses words of a sentence at a time to vectorize into vector. For this purpose, the n-gram parameter is set to be 2.

Train-Test Split

Data Set

Out of all observations 80% are used to train the model and 25% is used to test the model.

Training Data Set

It has 3045 observation and used to train the model

75 %

Train and Test Data having **N-Gram = 1**

Train and Test Data having **N-Gram = 2**

Testing Data Set

It has 1041 observations and used to test the model performance

25 %

Models

MODELS	MonoGram	BiGram	BestModel
Support Vector Classifier	Train: 99.8% Test: 99.01%	Train: 96.91% Test: 96.09%	MonoGram
K-Nearest Neighbour	Train: 68.6% Test: 68.2%	Train: 65.6% Test: 63.6%	MonoGram
Logistic Regression Classifier	Train: 96.6% Test: 94.1%	Train: 92.25% Test: 81.13%	MonoGram

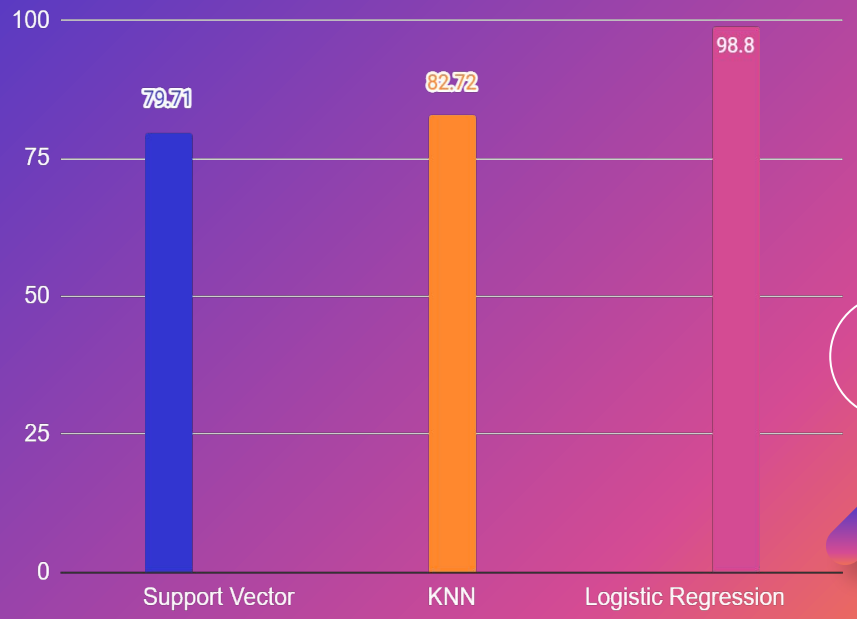
Evaluation

ROC -AUC Matrix

AUC - ROC score is a area under the curve of False Positive Rate and True Positive Rate.

The Roc score is maximum for Logistic Regression model although Support Vector Classifier gives the best possible accuracy.

Hence **Support Vector Classifier** is not preferred to use and we have a **Logistic Regression** model with the accuracy of 98.8%



ROC - AUC Score

Thank You