**SENTIMENT ANALYSIS ON PRODUCT REVIEW DATA**

**Ashwin Nair**

*Abstract- Under the current advances, the entire globe is rapidly changing. With the Internet*

*being used in every profession, it has become a basic demand for everyone. People are adopting social network applications to voice their perspectives on daily concerns, thanks to the increasing growth of these platforms. It is critical to collect and analyze people's reactions to purchasing a product, using public services, and so on. Sentiment analysis (also known as opinion mining) is a typical debate preparation assignment that seeks to uncover the sentiments that underpin opinions in texts on a variety of topics. In recent years, sentiment analysis experts have focused on assessing opinions on a variety of topics, including movies, commercial products, and everyday societal challenges. Amazon is the largest online platform for purchasing products and hence it has the largest set of reviews which can be used to detect the sentiments of the customers from the content of those reviews. As such our project explores the various sentiment analysis applied to user reviews on Amazon.*

**Introduction**

Sentiment analysis, also known as "opinion mining" or "emotion Artificial Intelligence," refers to the systematic recognition, extraction, evaluation, and examination of emotional states and subjective information using natural language processing (NLP), text mining, computational linguistics, and bio measurements. Sentiment analysis is concerned with the voice in client materials, such as surveys and reviews on the Internet and social media sites.

In general, sentiment analysis tries to figure out how a speaker, author, or other subject feels about a particular topic via positive or negative responses to a product, service, or occasion. Client surveys or recommendations on a wide range of topics are now widely available on the Internet, and audits may include surveys on topics such as clients or film fault-finding, among other things. Surveys are fast growing in popularity since people want to express themselves online. As a result, mining this data, differentiating client evaluations, and arranging them is a critical task. Sentiment mining is a task that employs natural language processing (NLP) and information extraction (IE) techniques to scan a large number of archives to compile the sentiments of various authors' remarks. Various methodologies, such as computational etymology and information retrieval (IR), are used in this procedure. The core concept behind sentiment analysis is to determine the polarity of text documents or short sentences and categorize them accordingly. The polarity of a sentiment can be classified as "positive” or "negative”. We have used Sentence-level sentiment analysis where each review or sentence can be classified as positive or negative.

**Definition and Motivation**

Sentiment analysis is a technique for analyzing the opinions of individuals or groups, such as a subset of a brand's followers or a single consumer in contact with a customer service agent. Sentiment analysis, as a scoring system, monitors conversations and evaluates speech and voice affectations to gauge emotions and feelings, particularly those related to a business, product or service, or topic.

Sentiment analysis is a method of determining if articulation is positive or negative in written or spoken languages. The current analysis technologies on the market are capable of consistently and precisely dealing with large volumes of customer feedback. Customers' thoughts on numerous topics, such as the purchase of things, the provision of services, or the presentation of promotions, are discovered using sentiment analysis in conjunction with content inquiry.

Surveys, online journals, comments, discourses, photos, and recordings are all used to offer massive amounts of client-created web-based social networking interactions regularly. These correspondences supply substantial opportunities to receive and comprehend consumer perspectives on topics such as fascination, as well as data that may be used to clarify and anticipate company and social news, such as product offers, stock returns, and the outcomes of political selections. The evaluation of the ideas transmitted by clients in their content exchanges is an important part of these investigations.

Amazon.com, Inc. is an e-commerce, cloud computing, digital streaming, and artificial intelligence-focused American multinational technology firm. There are 300 million active Amazon users. Every month, 197 million individuals visit Amazon.com. In February 2019, Amazon.com received 2.24 billion visits. Nearly two-thirds of Americans have purchased anything from Amazon (that's more than 90% of all online buyers in the United States!). In 2019, Amazon's net sales revenue was $280.5 billion. In five product categories, Amazon has a market share of more than 90%. Sentiment Analysis thus is very crucial for such a big platform where users hand over their trust to a 3rd party seller hence analysis of reviews may give a good idea of whether to purchase a certain product from a certain seller or not.

**Importance and Background**

Opinions are crucial to every single human activity since they are the primary determinants of our actions. We need to know what others are thinking at any point when we have to make a decision. In truth, businesses and organizations need to know what their customers think about their products and services regularly. This type of connection creates a huge window of opportunity for advertising information. People of every nationality, sexual orientation, race, and social class use the internet to share their experiences and impressions about almost every aspect of their lives. Opinion investigation is a type of data mining that solves these problems by meticulously extracting and examining web-based data without causing delays. Advertisers can continually discover buyers' feelings and states of mind with conclusion examination, despite the challenges of information structure and volume. There are two reasons for this study's enthusiasm about using sentiment analysis as a tool for encouraging research. Sentiment analysis is a key tool for businesses to determine their customers' preferences for items and brand image. It also plays an important part in evaluating data from various businesses and organizations in order to assist them in making business decisions.

**Classification Techniques**

Classification algorithms have been created in the field of machine learning, which employs various methodologies to classify unlabeled data. It's possible that classifiers will require training data. We have used the Naive Bayes classification, training a classifier effectively will make future predictions easier.

**Naïve Bayes**

This is a classification method based on Bayes' Theorem and strong (naive) independence assumptions between features. The closeness of a given feature (element) in a class is expected to be detached from the closeness of other elements by a Naive Bayes classifier. For example, an organic fruit that is red in color, spherical in shape, and is around three inches in width may be labeled an apple. Due to the likelihood that this natural fruit is an apple, a Naive Bayes classifier would consider these attributes independent, regardless of whether they are reliant on one another or on the presence of additional variables. Alongside effortlessness. The Bayes hypothesis is a method of computing for distinguishing likelihood P(a|b) from P(a), P(b), and P(b|a) as follows:



Where is the posterior probability of class a given predictor b and is the likelihood that is the probability of predictor b given class a. The prior probability of class a is denoted as p(a), and the prior probability of predictor p is denoted as p(b).

The Naive Bayes is widely used in the task of classifying texts into multiple classes and was recently utilized for sentiment analysis classification.

**Procedure**

**DATA PREPARATION**

For this project, I am using a dataset of amazon reviews. The dataset consists of 3.6M text reviews and their labels, I have used only a small fraction of data of about 10000. To prepare the dataset, downloaded data is loaded into a pandas data frame containing two columns – text and label. Next, I divided the dataset into two subsets, a training set of 8000 entries named trainDF and a testing set of 2000 entries named testDF. trainDF serves the purpose of training our classification model and testDF is used to check the accuracy of the model. I have created a custom encoder to convert categorical target labels to numerical form, i.e. (0 and 1). All the texts with the label “\_\_label\_\_1” are labeled as 0 and “\_\_label\_\_2” are labeled as 1.

**Pre-Processing**

Then I moved to the pre-processing stage where I first had to use the word lemmatizer on the contents of our corpus. Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form. This was done with the help of the nltk library which is a suite of libraries and programs for symbolic and statistical Natural Language Processing (NLP) for English written in the Python language.

**TEXT TRANSFORMATION**

**Bag of Words technique**

Bag of Words Model(BOW), which is used to represent the text in the form of a bag of words,i.e. the grammar and the order of words in a sentence is not given any importance, instead, multiplicity,i.e. (the number of times a word occurs in a document) is the main point of concern.

**The SciKit library** in python provides various methods to implement this technique. I have used tfidf-Vectoriser in this project since tf idf weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus and thus provides better and more accurate results.

Next, I have converted the text data into vectors, by fitting and transforming the corpora that I have created. I have made use of the n-gram parameter to give importance to words that have a very different meaning if used together. For eg: the word ‘social media’ has a very different meaning from the individual words ‘social’ and

'media’. I have set the value of the n-gram parameter to (1,2) which signifies a bigram (2 words at a time.)

**NAIVE BAYES CLASSIFIER ( MULTINOMIALNB )**

**Training the Classification model**

Now I have our dataset split, processed, and transformed into vectors ready to be used for training our classification model.

I have made use of the multinomialNB() classifier imported from the sci kit library. The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification).

The classifier is trained using the trainDF dataset and takes two inputs: texts and their corresponding labels.

**Testing the Model**

Before testing , the testing dataset testDF is preprocessed and transformed into vectors just as the training dataset. The prediction is evaluated using the predict function of the nb classifier. This prediction value further helps in calculating the accuracy score, precision and recall score of our classification model

Accuracy: takes two inputs, the labels of testDF and corresponding predictions made by the model and returns the accuracy of our trained model.

Precision Score : takes two inputs, the labels of testDF and corresponding predictions made by the model and returns the precision score of our trained model.

Recall Score : takes two inputs, the labels of testDF and corresponding predictions made by the model and returns the recall score of our trained model.

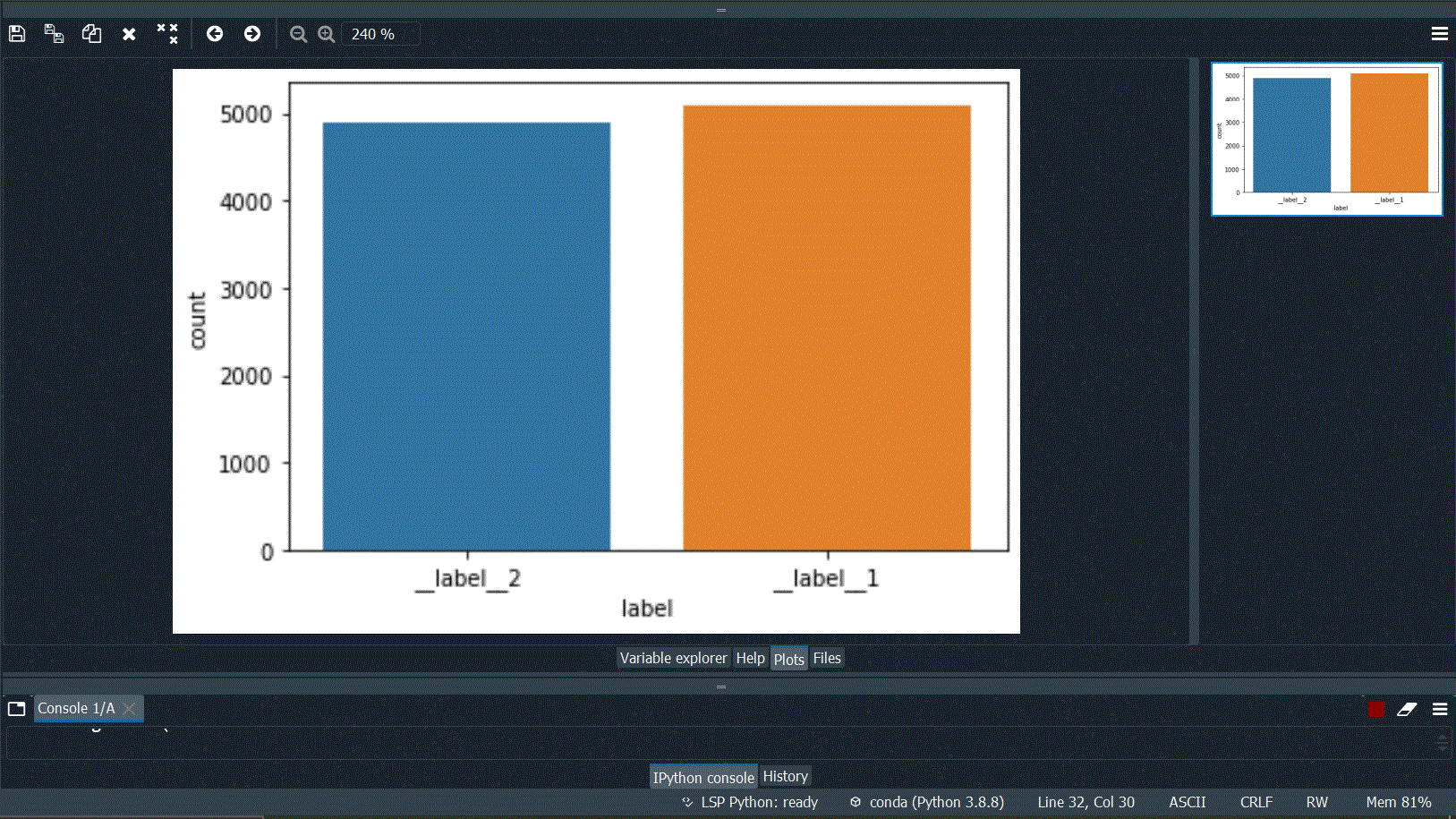
**Checking using custom input (The Sentiment predictor)**

I have also made a function to check the accuracy of our trained model using custom inputs. The sentiment predictor function takes the custom input and does all the preprocessing and transformation of the input into vectors to feed to the model . The prediction made by the model then goes into another function named ‘ Expression check’ which checks the prediction of the model and gives the final statement. If the prediction by the sentiment predictor is 0 then it prints it as a positive sentiment and if the prediction is 1, It prints it as a negative sentiment.

**DATA VISUALISATION**

**Label Count**

The seaborn library in python is used for various data visualization purposes. With the help of this library, I was able to plot a graph of the count of the reviews according to their labels.

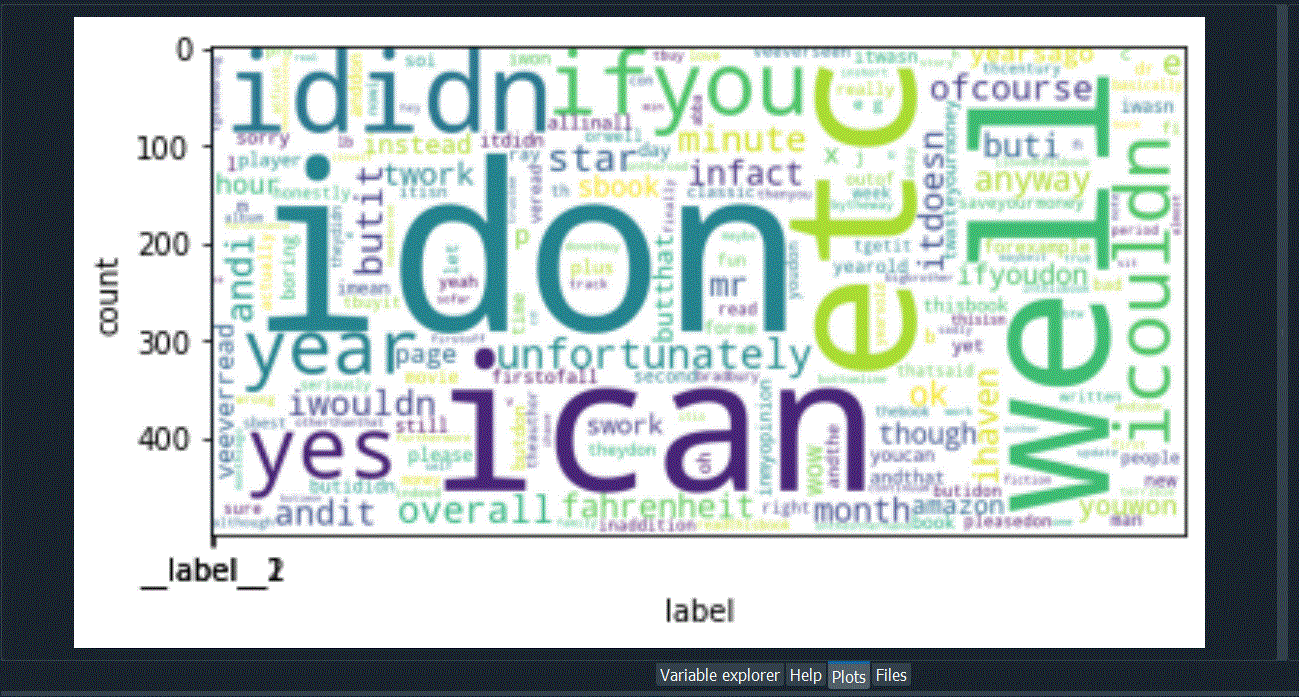


\_\_label\_\_2 (positive reviews) count = 4999

\_\_label\_\_1(negative reviews) count = 5000

**Wordcloud**

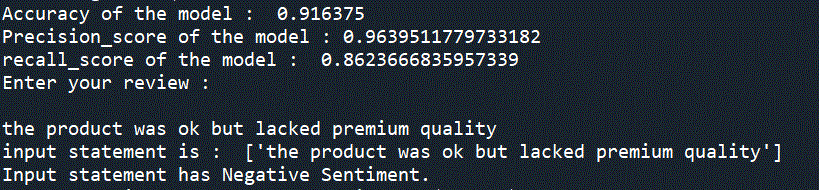
Another python library known as Wordcloud is very useful to identify the type of words contained in the dataset. It shows us the most frequently used words highlighting the importance of those words in the dataset which is very helpful in analyzing the tf-idf weights of the words contained.



This cloud shows us the words which are most frequently used in the dataset with a bigger font than those which are not used as frequently.

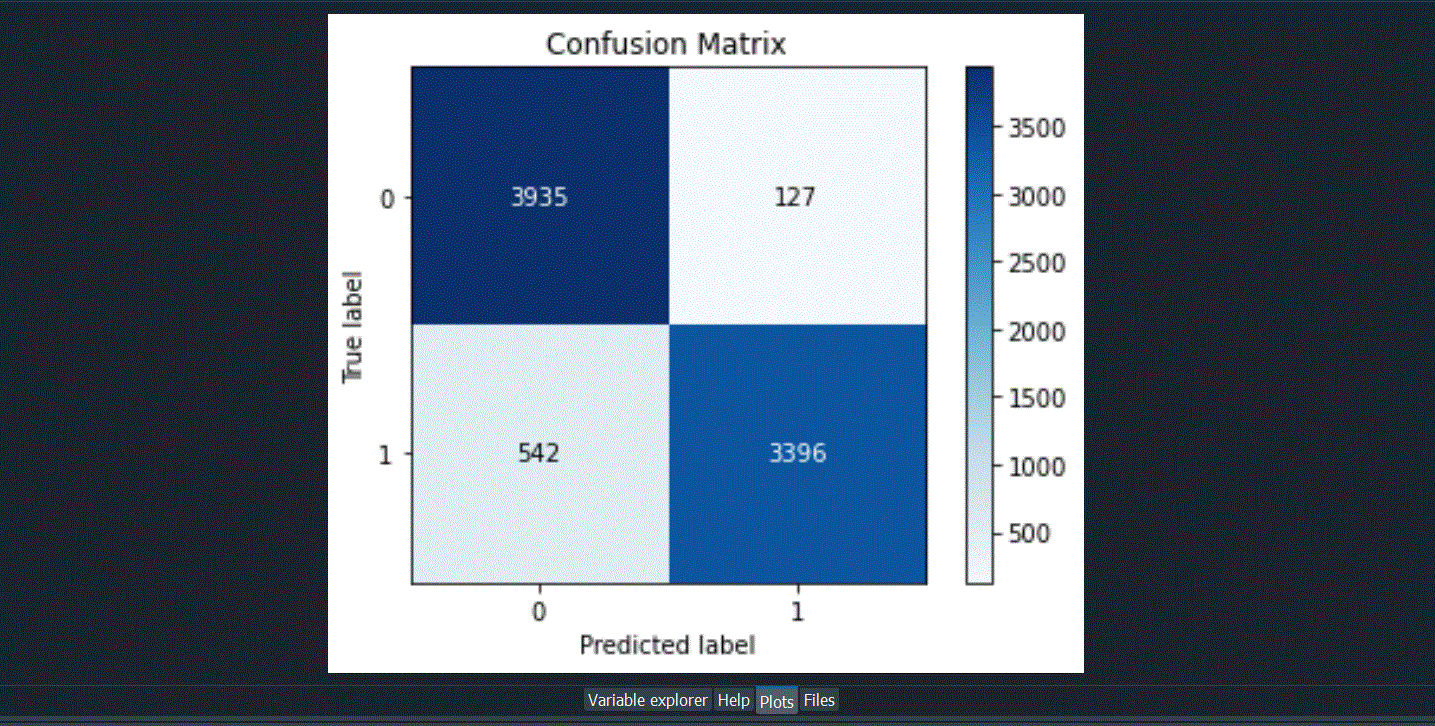
**Experimentation and Results**

I put our trained model to test and found out that our model had an accuracy of 91% with a precision score of 96% and a recall score of 86%.



To visualize the performance of our trained model, I made use of a confusion matrix,

Using a python library matplotlib, which shows the number of labels predicted correctly(True positives and True negatives) along with the number of labels predicted incorrectly ( False positives and False negatives).



**Flowchart**

