SENTIMENT ANALYSIS OF AMAZON REVIEWS USING MACHINE LEARNING MODEL

OMAR MOHAMMED ALI ALBAAGARI

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 4

INITIAL RESULTS

4.1 Overview

This chapter discusses the preliminary results and sentiment analysis of the Amazon office products. This chapter starts with the identification of the data set which is the exploratory data analysis (EDA), followed by VADER (Valence Aware Dictionary and sEntiment Reasoner) and Reberta (A Robustly Optimized BERT Pretraining Approach) analysis sentiment analysis, the creating and implementing the model using machine learning techniques.

4.2 Exploratory Data Analysis

EDA, which stands for exploratory data analysis, is one of the strategies that may be used to learn every information about the dataset. The patterns, trends, and relationships that are present within the data are included here. Before beginning to construct the machine learning method, it is quite helpful to have a solid grasp of the data structure already in place.

The text column describes the buyer reviewer of the product on Amazon platform in the office products category. Then the reviews will be analysed to obtain the results of sentiment analysis of the office products whether it is positive, negative or neutral.

One example of the raw dataset that was loaded from Python is shown in Figures 4.1 and 4.2, which can be found as follows. In all, there are 200000 rows of data, and there are 10 columns.

parent_asin	user_id	helpful_vote	asin	text	timestamp	images	verified_purchase	title	rating
B095CPWNTQ	AFITLXUBYKIELXW4EEA7IT5KEQQQ	0	B007D930YO	Was easy to setup and it work good with google...	1525408368407	[]	True	Was easy to setup and it work good with google...	5.0
B00BUV7C9A	AFITLXUBYKIELXW4EEA7IT5KEQQQ	0	B00BUV7C9A	Works as good with GV and works as good as Oom...	1441344761000	[]	True	Five Stars	5.0
B00006IEI4	AEB2U6KK3TFESGJY2PAHYW3M2QAQ	0	B00006IEI4	Works great!	1477875005000	[]	True	Five Stars	5.0
1604189274	AEB2U6KK3TFESGJY2PAHYW3M2QAQ	0	1604189274	Great quality, has many different uses.	1477874895000	[]	True	Five Stars	5.0
B09B1PNJ9Q	AEB2U6KK3TFESGJY2PAHYW3M2QAQ	0	B003F189HM	Love putting my students birthdays on these!	1477874814000	[]	True	Five Stars	5.0

Figure 4.1 Dataset

timestamp	images	verified_purchase	title	rating
1677939345945	[]	True	Pretty & I love it!	5.0
1677939160682	[]	True	2 excellent 1 extremely dry (blue)	4.0
1660188831933	[]	True	I don't get the reviews. Mine are garbage.	1.0
1659806066713	[{'small_image_url': 'https://m.media-amazon.c...	True	Ordering Ink online: never a good idea I guess.	4.0
1659799390978	[]	True	Mine are iffy at best.	3.0

Figure 4.2 Dataset

In figure 4.3 shows the dataset information of each column also the type of the data that used. It can be seen that all columns are non-null, consisting of 6 objects, 3 int64, 1 bool, and 1 float64.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 10 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   parent_asin        200000 non-null  object
 1   user_id            200000 non-null  object
 2   helpful_vote       200000 non-null  int64
 3   asin               200000 non-null  object
 4   text               199975 non-null  object
 5   timestamp          200000 non-null  int64
 6   images             200000 non-null  object
 7   verified_purchase  200000 non-null  bool
 8   title              199964 non-null  object
 9   rating             200000 non-null  float64
dtypes: bool(1), float64(1), int64(2), object(6)
memory usage: 13.9+ MB
```

```
df.columns
```

```
Index(['parent_asin', 'user_id', 'helpful_vote', 'asin', 'text', 'timestamp',
       'images', 'verified_purchase', 'title', 'rating'],
      dtype='object')
```

Figure 4.3 Dataset Information

In figure 4.4 shows the dataset description which is about the basic statistical analysis such as mean, standard deviation, minimum and maximum values.

```
df.describe()
```

	helpful_vote	timestamp	rating
count	200000.000000	2.000000e+05	200000.000000
mean	1.108975	1.545500e+12	4.412790
std	10.618726	8.653161e+10	1.111684
min	0.000000	9.587741e+11	1.000000
25%	0.000000	1.482169e+12	4.000000
50%	0.000000	1.558833e+12	5.000000
75%	0.000000	1.614727e+12	5.000000
max	1561.000000	1.679245e+12	5.000000

Figure 4.4 Dataset Description

A word cloud including reviews with positive sentiments was shown in the figure below. Through the use of word cloud analysis, it was determined that the words "easy," "perfect," "love," "great," "nice," "beautiful," "good", "pen", "better" were the

ones that were used the most often in the evaluation of the products. It was clear that the office product was helping them to meet with their needs since the term "good quality" was used. The term "works great" was a representation of the products that satisfy them. During this time, the words "price" and "happy" were used to symbolize their satisfaction of the price of the products which is affordable.

Figure 4.5 World Cloud of Positive Sentiment

A word cloud including reviews with negative sentiments was shown in the figure below. Through the use of word cloud analysis, it was determined that the words "small," "little," "bad," "waste," and "disappointed" were the ones that were used the most often in the reviews of the products. The words "time," "hard," and "small " demonstrate that a significant number of individuals will have a negative attitude towards a number of the office supplies that have been purchased.

Figure 4.6 World Cloud of Negative Sentiment

Figure below illustrate the word cloud for neutral sentiment reviews. Word cloud analysis illustrated that "use," "color," "wish," "product," were the most frequent used words in the review. These words can represent that the products don’t meet their meet as their expectation.

Figure 4.7 World Cloud of Neutral Sentiment

A representation of the rating distribution for the office product through amazon seen in Figure 4.8. The rating ranges from one to five, and as can be seen, the majority of the ratings are five, which indicates that the majority of sellers are offering items of a high quality at prices that are within reasonable ranges. Indicating most of the customers are satisfied with their purchase.

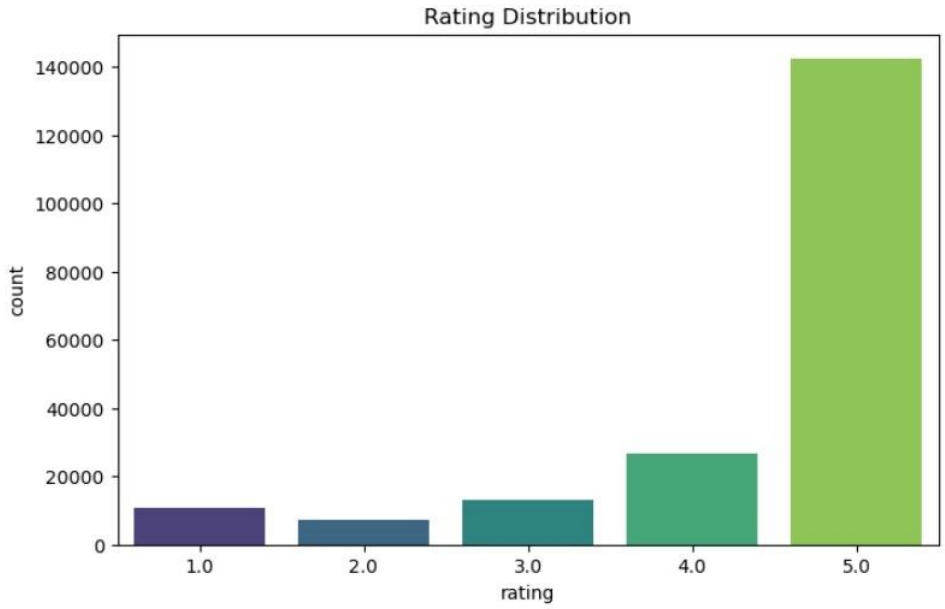

Figure 4.8 Rating Distribution

Figure 4.9 illustrate the verified purchase distribution which indicate the majority of the customers are verified their purchase. Whereas smaller portion are not verifying their purchase.

Figure 4.9 Verifies Purchase Distribution

A list of the top ten titles of reviews submitted by consumers is shown in the figure below. "Good Product" accounts for the smallest share, while "Five Star" is the term that receives the most amount of portion.

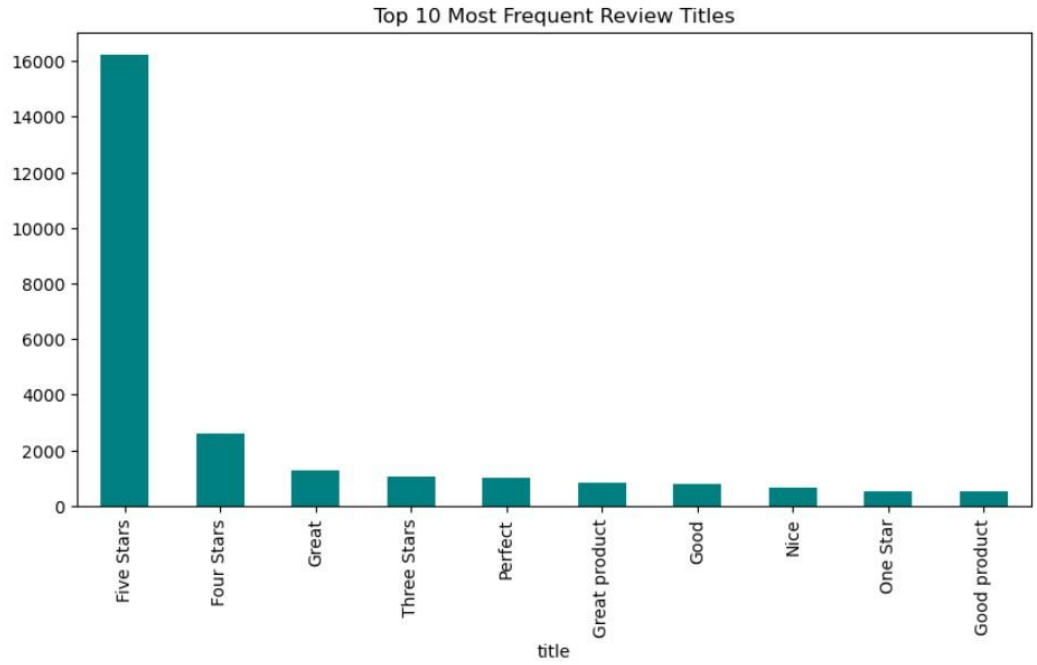

Figure 4.10 Top 10 Most Frequent Review Titles

A helpful association between rating and helpful vote was indicated by the scatter plot as illustrated below. The growing number of customers who considered the evaluations to be helpful, as well as the rising ratings of the ability of the products meet their needs. As a result, helpful vote that have received higher ratings are more likely to the customer shows that the product is helpful.

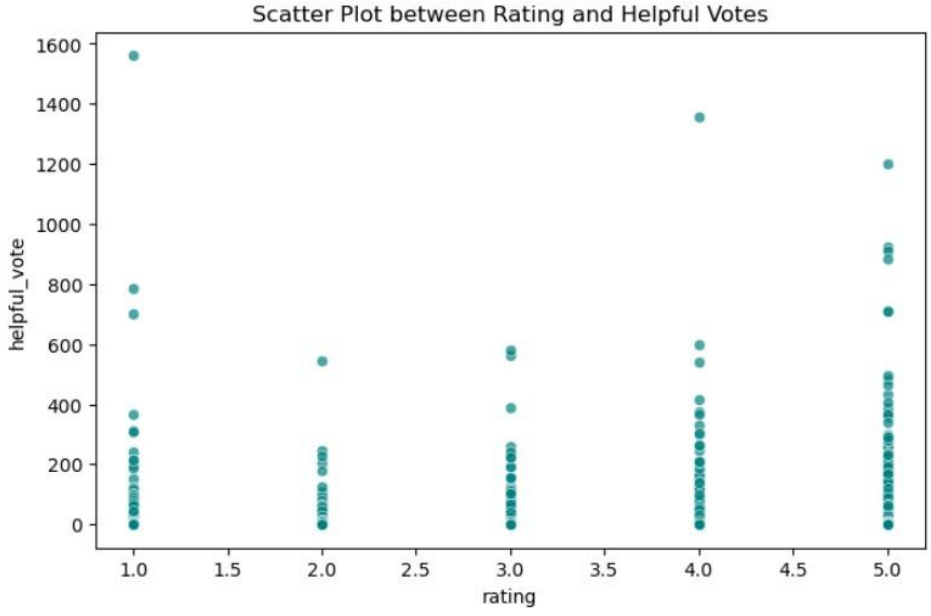

Figure 4.11 Scatter Plot between Rating and Helpful_vote

4.3 Data Cleaning

During the process of sentiment analysis, data cleaning is an essential step, particularly for the purpose of ensuring that the data utilized is clean, pertinent, and capable of being processed effectively by the model. The following are the processes that were taken to clean the data that was collected from the Amazon dataset of the office products category. Figure 4.12 illustrate the steps of the data cleaning.

Removing the unnecessarily columns help to minimize the columns in the dataset to work with in easily. Moreover, rearranging the columns in a specific way to make easy to reach. Furthermore, drop any rows that having missing values, convert to lower case which the majority of the letters in the text have been changed to lowercase. Due to the fact that both capital and lowercase letters are handled in the same manner, the analysis becomes more consistent. For instance, the terms "LIKE" and "like" are often used interchangeably. lastly, remove digits, punctuate, and extra space to make the text more effective and neater to do the sentiment analysis.

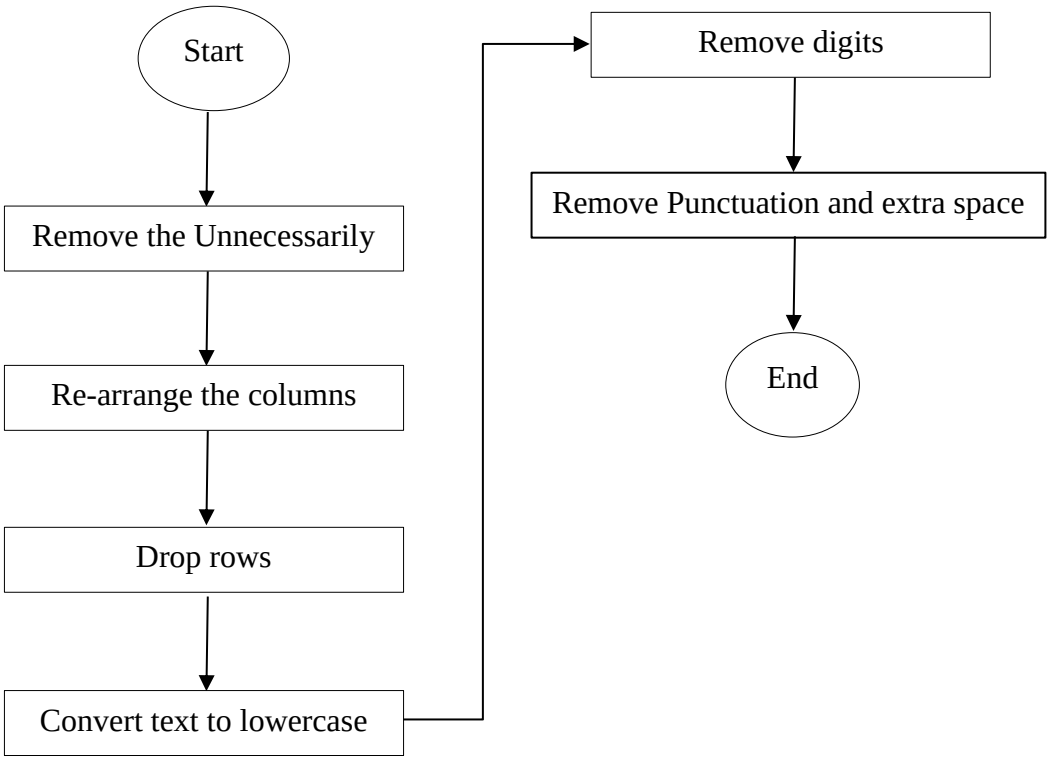

Figure 4.12 flowchart of data cleaning

4.4 Data Preparation

In order to guarantee that the data is clean and formatted before any further processing takes place, the preparation of the data is an essential step. Therefore, the reviews by the customer in the dataset should apply to them snowball stemmed to return the word to it's original like 'lovely' become 'love' to make the process faster later on when applying the sentiment analysis

Therefore, tokenize the text and slice it for faster processing, this technique is known as tokenisation, and it involves separating a text into its component words, also known as tokens as illustrated in the figure below.

```
example = df_cleaned ['Text'] [49]
print (example)
```

```
This helps you find the remotes.
```

```
tokens= nltk.word_tokenize (example)
tokens[:10]
```

```
['This', 'helps', 'you', 'find', 'the', 'remotes', '.']
```

Figure 4.13 Example of Tekenisation

The next phase is Part-of-speech (POS) tagging, which is used to give a part-of-speech tag to each tokenized word in the review. POS tagging plays a significant role in numerous NLP tasks by offering linguistic insights and enabling the analysis and interpretation of textual data.

```
tagged= nltk.pos_tag(tokens)
tagged[:10]
```

```
[('This', 'DT'),
 ('helps', 'VBZ'),
 ('you', 'PRP'),
 ('find', 'VBP'),
 ('the', 'DT'),
 ('remotes', 'NNS'),
```

Figure 4.14 POS Tagging

4.5 Sentiment Analysis

Within the context of this section on sentiment analysis, we will locate each word that appears in reviews from Amazon dataset and classify it according to whether it is positive, negative, or neutral according to Valence Aware Dictionary and sEntiment Reasoner and Reberta analysis. Take a look at the following examples of sentences that are positive, negative, and neutral.

Table 4.1 Some Examples of Sentiment Analysis sentences

Review	Sentiment
They work great. I love the colours	Positive
These were okay some worked and some didn't No worries don't have this printer anymore	Neutral
Not impressed. Print is very poor quality. Looks faded.	Negative

4.6 Model Development

In this research using two models in the sentiment analysis which are Valence Aware Dictionary and sEntiment Reasoner and Robustly Optimized BERT Pretraining Approach using Python. Some libraries were used including the NLTK, vaderSentiment, Transformers, Pandas, Scikit-learn, and seaborn. Those are helping to approach the sentiment analysis in effective way.

For the rule-based sentiment analysis, the VADER (Valence Aware Dictionary and sEntiment Reasoner) model was implemented using the NLTK library. VADER operates by assigning sentiment scores to text based on a pre-defined lexicon and set of rules. It produces four scores: positive, negative, neutral, and a compound score that combines these sentiments into a single metric .Sentiment classification was performed by applying thresholds to the compound score, categorizing reviews as positive, neutral, or negative. Due to its simplicity and efficiency, VADER was particularly useful for quickly analysing large volumes of reviews.

The phrases "neg," "neu," and "pos" are used in VADER data frame as shown below is referring to various characteristics of the sentiment that is communicated in a piece of text. These elements are influenced by the degree to which the text contains positive, negative, and neutral feelings, and the strength of those sentiments.

	Id	neg	neu	pos	compound	Product_ID	Helpful_Vote	Rating	Time	verified_purchase	Summary	Text
0	1	0.000	0.677	0.323	0.9300	B01MZ3SD2X	0	5.0	1677939345945	True	Pretty & I love it!	Lovely ink. Writes well. The right amount of w...
1	2	0.051	0.771	0.178	0.9481	B08L6H23JZ	0	4.0	1677939160682	True	2 excellent 1 extremely dry (blue)	Overall I'm pretty happy with this purchase bc...
2	3	0.070	0.815	0.115	0.9498	B07JDZ5J46	2	1.0	1660188831933	True	I don't get the reviews. Mine are garbage.	[[VIDEOID:63276c19932aa4f3687042b8b9f8613c]] U...
3	4	0.072	0.755	0.173	0.9941	B07BR2PBJN	0	4.0	1659806066713	True	Ordering Ink online: never a good idea I guess.	It's a beautiful color, but even though it had...
4	5	0.142	0.776	0.082	-0.9306	B097SFY5ZS	0	3.0	1659799390978	True	Mine are iffy at best.	Idk if I just got a bad batch which is possibl...

Figure 4.15 Data Frame of Vader Model

The neg score, also known as the negative score, is a statistic that represents the percentage of negative emotion that is present in the text. The amount to which the language displays negative emotions, such as despair, or frustration, is represented by this metric. The value neg can range from 0 to 1, with higher values signifying a stronger degree of negativity than lower values on the scale.

The pos score, also known as the positive score, is measurement of the percentage of positive emotion that is present in the text. The amount to which the text communicates pleasant feelings, such as happiness, joy or contentment, is represented by this metric. Additionally, the value pos can vary from 0 to 1, with larger values suggesting a more robust positive.

The neu score, also known as the neutral score, is a type of score that reflects the percentage of the text that has neutral sentiment. The extent to which the text is free of strong emotional polarity or lacks a strong emotional polarity is represented by this. A higher neu score indicates that the text comprises a greater amount of neutral

language and a lower amount of information that is emotive. As is the case with neg and pos, the value neu might be anywhere between 0 and 1.

In the context of a piece of writing, the compound score is a single value that indicates the overall sentiment polarity of the text. In order to offer a thorough evaluation of the text's sentiment, it takes into account not only the positive and negative sentiment ratings, but also the intensities of those values. Between -1 and 1, it is a range.

This suggests that the majority of the reviews are positive and have a strong sentiment polarity in the positive direction, whilst the number of negative reviews is quite minimal and has a low sentiment polarity.

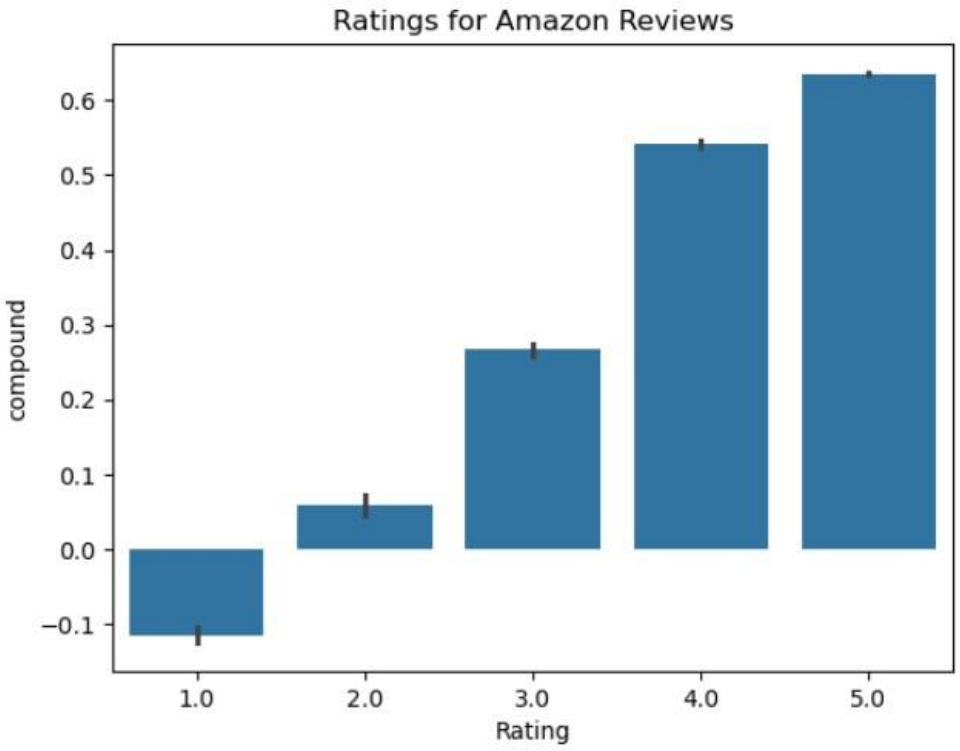

Figure 4.16 Visualization of VADER Sentiment

In contrast, the RoBERTa (Robustly Optimized BERT Pretraining Approach) model provided a more sophisticated approach to sentiment analysis. RoBERTa is a transformer-based deep learning model capable of understanding contextual

relationships in text. A pre-trained RoBERTa model from the Hugging Face library was utilized, specifically designed for sentiment analysis tasks. The text data was tokenized, padded, and truncated to meet the model's input requirements. If fine-tuning was conducted, the model was trained further on the Amazon reviews dataset to improve its performance. RoBERTa then classified the sentiment of each review based on the learned patterns in the text.

The phrases "neg," "neu," and "pos" are used in RoBERTa data frame as shown below is referring to various characteristics of the sentiment that is communicated in a piece of text. These elements are influenced by the degree to which the text contains positive, negative, and neutral feelings, respectively, and the strength of those sentiments.

roberta_neg	roberta_neu	roberta_pos	Product_ID	Helpful_Vote	Rating	Time	verified_purchase	Summary	Text
0.001184	0.016531	0.982284	B01MZ3SD2X	0	5.0	1677939345945	True	Pretty & I love it!	Lovely ink. Writes well. The right amount wet/...
0.066594	0.202709	0.730696	B08L6H23JZ	0	4.0	1677939160682	True	2 excellent 1 extremely dry (blue)	Overall I'm pretty happy purchase bc ink good ...
0.907260	0.081818	0.010923	B07JDZ5J46	2	1.0	1660188831933	True	I don't get reviews. Mine garbage.	[[VIDEOID:63276c19932aa4f3687042b8b9f8613c]] U...
0.156052	0.423542	0.420406	B07BR2PBJN	0	4.0	1659806066713	True	Ordering Ink online: never good idea I guess.	It's beautiful color, even though packed extre...
0.745801	0.219413	0.034786	B097SFY5ZS	0	3.0	1659799390978	True	Mine iffy best.	Idk I got bad batch possible I suppose bc let'...

Figure 4.17 Data frame of Roberta Model

4.7 Summary

In conclusion there were two models were developed in this work which are VADER and Roberta to do the sentiment analysis of office product reviews in Amazon. The VADER sentiment is provided to give some insight of the behavior of the customer as positive, negative, and neutral. The most portion of VADER analysis is positive which reflect that the customer are satisfied. On the other hands, Roberta analysis was developed by the data frame.