Major_Project_Report.pdf

by

Submission date: 21-Apr-2020 09:21PM (UTC+0530)
Submission ID: 1303697521
File name: Major_Project_Report.pdf (2.7M)
Word count: 7460
Character count: 40157

ABSTRACT

One of the most difficult and hard problems that a lot of e-commerce sites face today is the presence of an abundance of fake reviews by malicious users. This is usually done to eradicate competition. Currently there are no datasets to find out the fake reviews in a flawless way. So, we intend to use supervised machine learning with the help of cloud computing to fabricate an efficient fake review analysis monitoring system. To achieve this goal, preprocessing of data must be done to ensure that there is no flaw in the input data. It is a very important step as any error in the data will ultimately cause an error in the analysis. Once the processing of data is done, the reviews are analyzed in several stages including Stemming, Bag of words method and Tokenization. These techniques facilitate in the detection of fake reviews. Online reviews have a pivotal part in online-shopping as the customers tend to see the reviews before deciding to buy any product and hence fake reviews are a major concern. So, it is very important to remove these Deceptive reviews that are posted on the e-commerce websites and help customers avoid falling for the trap by showing them only the genuine reviews.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In today's world, data is going through an exponential growth. E-commerce sites such as amazon receive large amounts of user generated data such as reviews. These reviews are an integral part of an e-commerce site as it makes or breaks the sales of a product. This is mainly due to the fact that customers look into the reviews before deciding on whether to buy the product or not. But as we all know there are people and organizations who exploit this feature by submitting fake reviews through the use of bots consequently affecting the sales of the product. These immoral actions are usually done by competitor's or hackers. In order to prevent this, we have decided to create a solution for this problem using python. The fake product review analysis systems before have only been done by data or opinion mining. But we have decided to use Random Forest Classifier instead.

CHAPTER 2

LITERATURE REVIEW

2.1 EXISTING SYSTEM ANALYSIS

Due to the several processes involved, the already existing systems take up more memory and time, than the optimum amount, to analyze the review which in turn affects the efficiency of the system. The current efficiency of the system does not keep up with the fast-paced world. Due to the lack of speed the fake reviews can only be viewed by the customer by the time it could have been detected and removed. Even though the fake reviews are only online for a small period of time it matters a lot in the e-commerce market.

2.2 OPINION MINING

Starting with the categorization on various types of reviews: opinion false reviews are of two types. Positive for upgrading the sales of product and negative for degrading the sales of the product, reviews on brands by the sole company to upgrade the product despite viewing the quality, non-reviews which are not related to product itself but posted by system generated software. [1] Detection of spams depending on the overall rating of the product, group of fake reviews in a continuous manner and online review manipulation in graphic format. [2] Starting with tracking an IP address of the reviewer, if multiple reviews are from the same source then it is considered to be fake. Reviews on brand only are also considered because the value of brand can't play a role in defining the genuinely of any product. [3] A pool of words which are categorized as positive for a real review and negative for a spam review, if such words are identified then the results are straightforward. Starting with the drawbacks of existing models, distinguishing between fake and real reviews. [4] User verification while writing a review which makes it easy for the admin to understand the polarity of the review. Here, the reviews are initially categorized into different forms. Namely, False reviews, Brand reviews and Spam review. Then, the deviations present in the ratings recorded are comprehended into a particular form. Then the sentiment analysis of the product review is done after IP address detected.

Firstly, the User IDs for the admins are verified. The reviews are then deleted or posted by the admins once the verification is done. [5] Finally, using the review polarity technique, the reviews are analyzed and then in the end, they are categorized as genuine or fake. Similar to the one previously mentioned, this paper also performs tracking of IP address initially. [6] Then, the accounts that were used to write the reviews are identified. Further, the reviews that are identified are divided into different types. Again, the main drawback of following this technique is consumption of time. [7] If not for the various steps present, the time taken which is high here could have been reduced by optimizing and reducing the number of modules present. Finally, the negative words are removed and the spam reviews are classified.

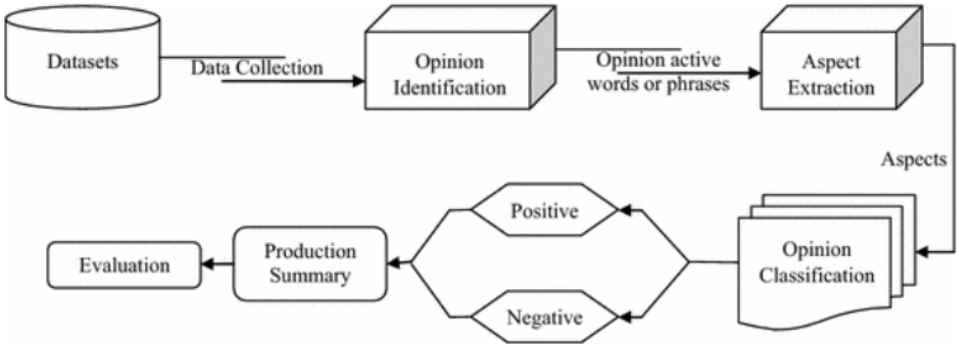


Figure 1:Opinion Mining

2.3 DATA MINING

Starting with research and survey and finding the drawbacks in the existing approaches, finding new strategies and solutions on how to extend in order to overcome the drawbacks. Creating a dataset from random records on available websites. [8] Combining the records collected into a coherent form. Separating the spam and non-spam reviews on the basis of preprocessing techniques like tokenization, stemming, etc. [9] Evaluating on the basis of approaches and conveying a final result. Survey conducted for the public to get a brief knowledge about the perspective of reviews on products. Starting with data cleaning using various programming languages. Sentiment analysis and prediction on different types of reviews. Fake review detection by performing machine learning algorithms like naive Bayes, linear regression. [10] Statistical analysis of reviews which makes the product valuable for the customers or not. Here, sentiment analysis has been used to analyzing of

the review. That is, the final outcome will be in the range of -1 to +1, a polarity wise categorization process. [11] But the specific nuances of the review cannot be handled using Sentiment Analysis making this method not optimally efficient in terms of determining if the review is fake or real. [12] Furthermore, Naïve Bayes and Decision Tree algorithms are also used. Here, the dirty dataset is firstly cleaned using various programming languages. The reviews are then analyzed using sentiment analysis. The analyzed reviews are further classified on the basis of genders. [13] The classified data is put in a testing classified and then the accuracy of the testing is determined. The processes performed are visualized neatly with the help of data visualization plots and graphs and in the end a statistical analysis is performed.

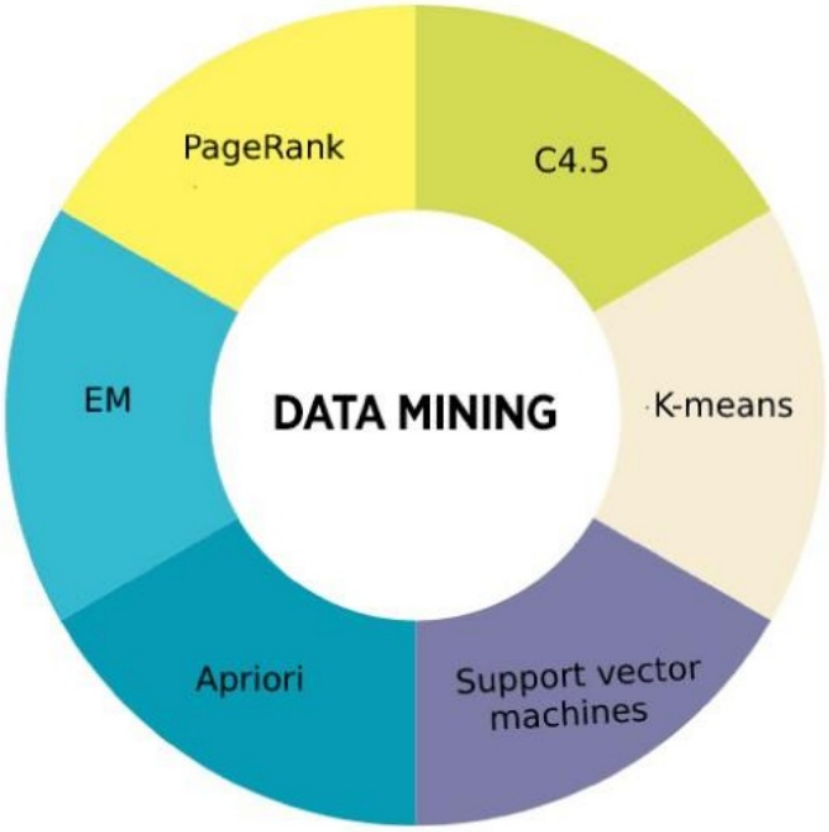


Figure 2:Data Mining

2.4 SENTIMENT ANALYSIS

Starting with the data mining process to scrutinize and store huge amounts of data and find different range of patterns related to the data. NLP for analyzing the opinions of the public depending on the text or numbers written on their reviews. Web scraper for scraping out the required content from the website. [14] Preprocessing techniques used to filter the reviews based on spam and non-spam category. Sentiment analysis for identification of fake reviews and content similarity for giving polarity to a particular review. Similar to the one previously mentioned, this paper also performs tracking of IP address initially. Then, the accounts that were used to write the reviews are identified. [15] Further, the reviews that are identified are divided into different types. Again, the main drawback of following this technique is consumption of time. If not for the various steps present, the time taken which is high here could have been reduced by optimizing and reducing the number of modules present. Finally, the negative words are removed and the spam reviews are classified.

Sentiment analysis is one of the difficult machine learning algorithms which detects the sentiment of texts and produce results (positive or negative opinion). The sentiments which are expressed by customers online can be real or fake, these sentiments can be traced out using this technique.

It provides an understanding of a customer's mind while buying a product depending on their reviews. The emotion of buying it varies from customer to customer and it affects the business.

Sentimental analysis work can be segmented into three categories. First category is to understand the polarity of the reviewer as positive, negative and zero. Second category focuses on emotions of the reviewers if they are happy or sad depending on their behaviour. Third category showcases the intentions of the reviewer whether they are interested or not in the product.

Here are some of the most popular types of sentiment analysis:

1. **Fine grained sentimental analysis**

This sentiment analysis type focuses on the first category as described above. It enhances the reviews on the basis of their polarity. Generally, they have limited categories to get a feedback but for getting a neutral feedback they divide the categories furthermore. It is arranged as the following category is expressed with an added condition. Positive feedback can be even more positive depending on the review and same goes for the negative one.

This way we can divide it into five categories:

- More positive
- Less positive
- More negative
- Less negative
- Neutral

Many products also include ratings where reviewers provide with number of stars with a scale from one-star to five-star ratings. While exhibiting the product there is an average method used to show the overall star rating given from various reviewers.

2. Emotion detection

This analysis showcases various emotions which reviewers express in their reviews. There are plenty emotion detection algorithms which uses their own ways to find the way of expressing emotions but the very famous technique is the use of Lexicons. It contains handful of words expressing their meanings. The drawback using this technique is language as a word used in the review can be a right word used for a wrong term. Reviewers express their emotions in certain way that it is difficult to understand whether they are praising the product or degrading its value. For example, one may consider a

review as "Killer", now this can be taken in a positive way even the words in it are negative.

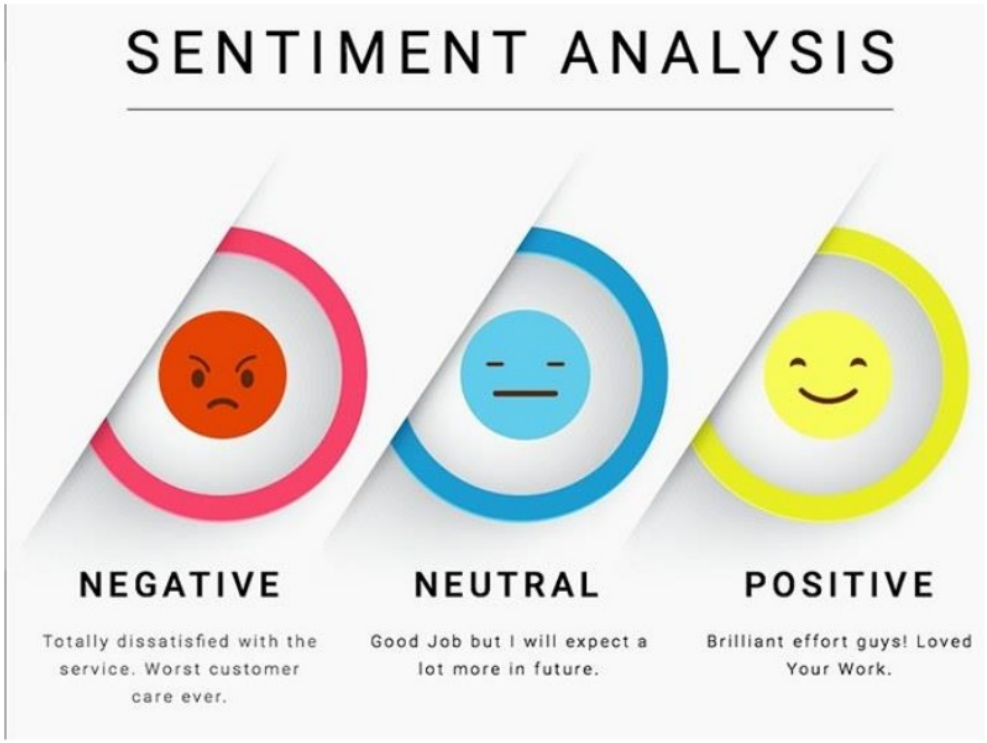


Figure 3:Sentiment Analysis

CHAPTER 3

PROBLEM DEFINITION

Product reviews are having a greater effect on the consumers than there used to be. Before buying a product almost everyone checks for star rating and reviews for the particular item. The average consumer wants to know if the product is good or not. Holding such high value, reviews play a crucial role in the online marketplace, potentially deciding if the product is sold or not. The downfall being they can be easily faked by competitors. The usage of bots is at a high and because of this, the buyer can easily be misled into picking the wrong choices. The practice of producing large volumes of fake reviews is called opinion spamming. The outcomes are often very polarising and aimed at being deceptive with either positive or negative reviews to either promote or demote a particular product. The people responsible for such acts are called fake or spam reviewers. With the increasing number of spam reviewers, the current system proves very ineffective in removing the fake ones as even a slight delay can cause severe damages to the product. And hence there is a great need to improve the system that the online system relies upon.

CHAPTER 4

ARCHITECTURE

4.1 PROPOSED SYSTEM ARCHITECTURE

All the programming work is done on a platform called Anaconda Navigator. Its GUI contains different various attributes as follows:

- Users.
- Apps, Dashboards, Notebooks.
- On-premises, Cloud are two deployment methods.
- SQL
- No SQL
- Cloud based storage

Above mentioned attributes are connected to Jupyter Notebook with a programming language called python. This notebook consists of various attributes as follows:

- User
- Browser
- Notebook servers consisting of notebook files
- Kernel

All these are interdependent on each other and everyone to work simultaneously for better outcomes in the Jupyter notebook.

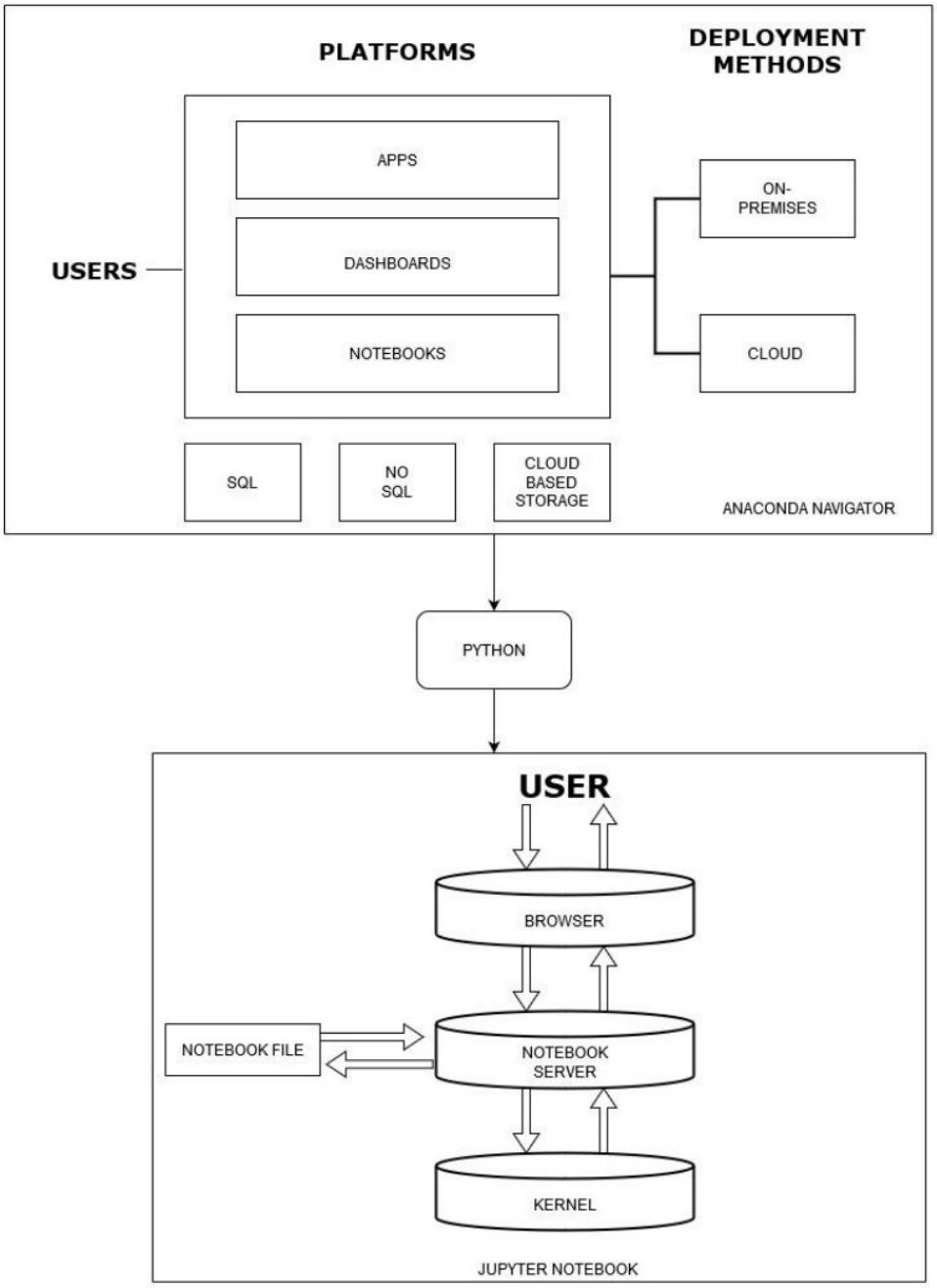


Figure 4:System Architecture

4.2 PROPOSED SYSTEM WORKFLOW

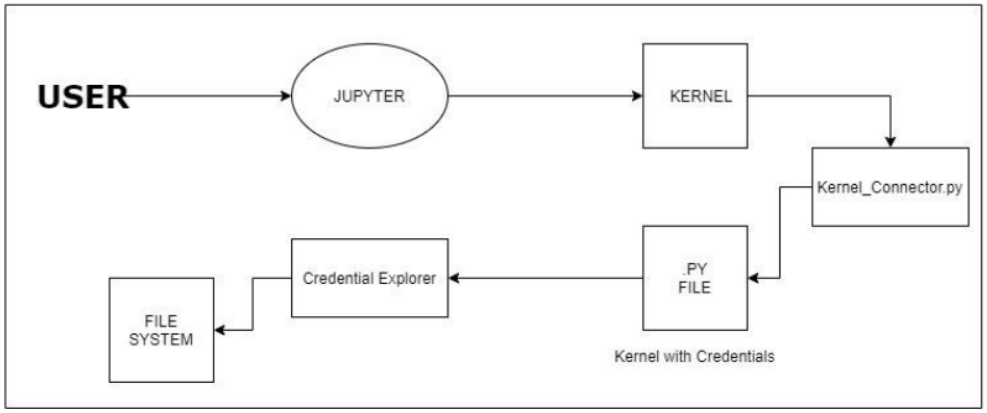


Figure 5:Data flow architecture

This shows the flow of data. Any general user can enter a Jupyter notebook having kernel as an operating system. Using python create a file named kernel_connector.py having all the credentials of the file. Have a credential explorer for connecting it to a file system where all the files to be maintained at one place.

4.3 FUNCTIONAL REQUIREMENTS

- **Functional Requirements:**

The system must take in product reviews from the database. It must identify and remove the fake reviews that are present.

- **Non-Functional Requirements:**

The raw data from the database must be pre-processed to suit the ML format and is later broken-down step by step.

- **Software:**

Python is used because of its robust standard library and the availability of open source frameworks and tools. In order to work in Python, Anaconda3 is environment is set up so that we can use the Jupyter notebook to write and test our program.

Figure 6:Python

Python

Python is a high-level programming language which was first used Thirty years ago. But it was not until recently that it was exploited to the maximum. Python in the recent past has had a great surge in the user count. This is because of its ease of usability and user-friendliness. With the inclusion of NumPy and Pandas, Python has become the go-to hub for Data Science. And it saw a quick jump of users from various other languages like MATLAB and Octave due to this reason. Users can upload packages of their own to python thus avoiding the need to run the same large codes repeatedly. Huge companies like Google have also made their contributions to packages. Keras and TensorFlow are increasingly being used by many. Such has been the impact of Python.

Figure 7:Anaconda

Anaconda Navigator

Anaconda navigator is a platform with which users can make use of various applications and techniques. This platform has several ways of doing it and it is made possible with the help of Dashboards, Apps and Notebooks. The most commonly used environments in Anaconda Navigator include the Jupyter notebook, Spyder Application and RStudio. Furthermore, header files and libraries for required languages can be inherently downloaded in the Anaconda Interface, thus avoiding the process of download packages from the internet or using command prompt. In addition to that, Anaconda Navigator also provides on-premises and cloud-based deployment methods.

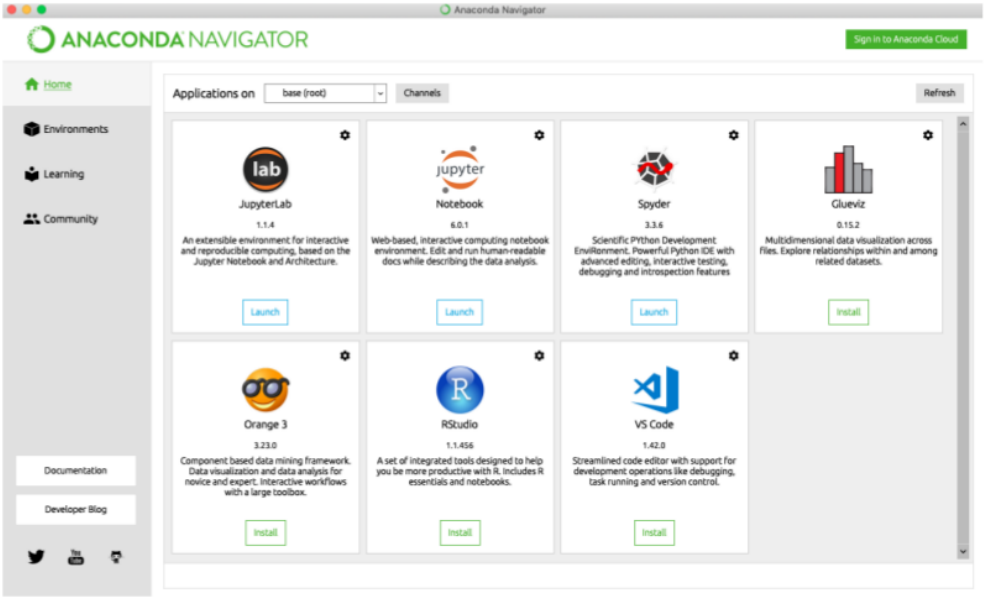


Figure 8: Anaconda Navigator

Figure 9: Jupyter

Jupyter Notebook

Jupyter Notebook is much more than an IDE. It is a web application that not only allows users to write code but also gives us a presentable format to explain it with. It houses the usage of Python and runs code with the help of cells. These cells run a single line of code, one at a time. Along with running code, comments can also be easily made with the help of

the markdown cells. These markdown cells are also used in writing in the Heading format. Namely, the H1 to the H6 tags. The .ipynb file is opened with the help of Jupyter notebook typically on a browser with the notebook server, which in turn is connected to the kernel. It also provides easy sharing and exporting options, as once we finish and run all codes, it can be exported to other formats like PDF and so on, thus making it very user-friendly.

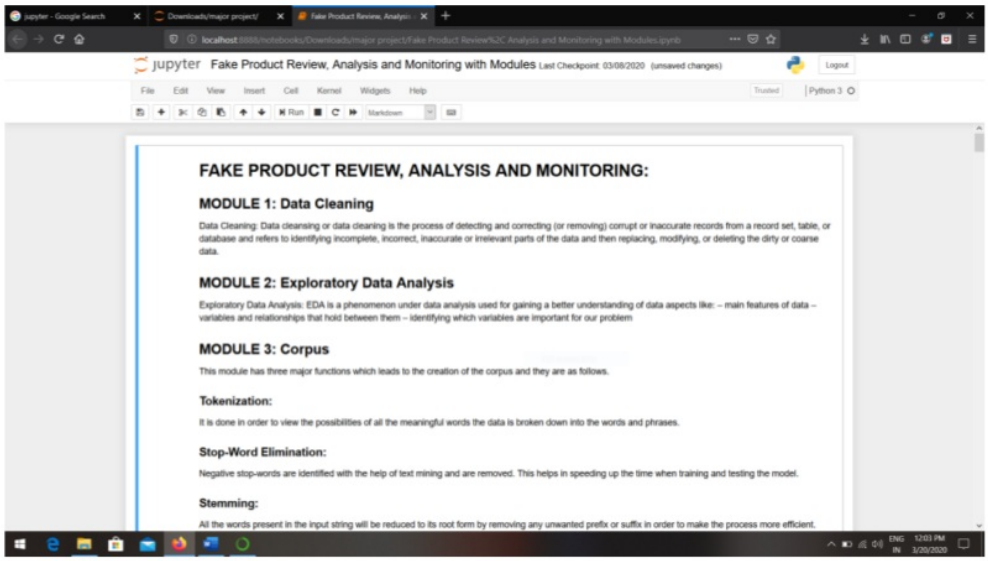


Figure 10:Jupyter Notebook

CHAPTER 5

PROPOSED METHODOLOGY

5.1 GENERAL OVERVIEW

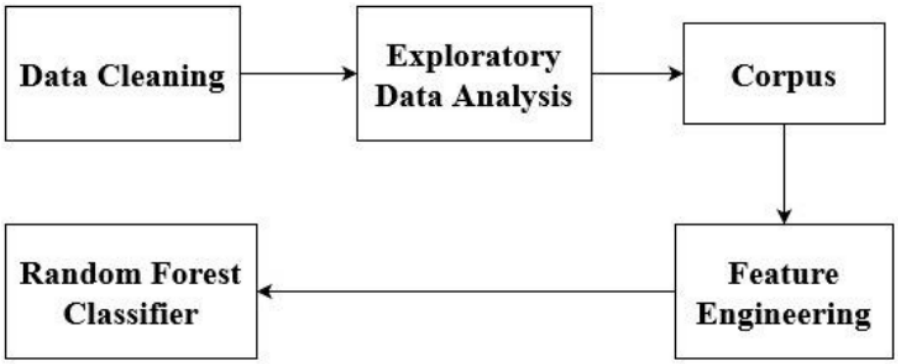


Figure 11:Proposed Methodology

STEP 1: Data Cleaning:

Data cleaning is the process where all the null or corrupted values in the csv/xlsx file we imported are modified or removed. These reviews are modified when the null values are huge in number or completely removed when the null values are only negligible in number.

STEP 2: Exploratory Data Analysis:

In this step, in order for us to know about the data we are handling visualization of the dataset is done using python visualization libraries.

STEP 3: Corpus:

Punctuation, Stop-words and stemming are performed in order to form a corpus which in turn is used in formation of bag of words.

STEP 4: Feature Engineering:

Feature engineering is a method of extracting of data and converting them into a format where the machine learning model can understand. Here, we can convert all the string values into numbers for faster training.

STEP 5: Random Forest Classifier:

In this step, the system is trained to spot fake reviews using a Random Forest Classifier Machine Learning Model. This is done with the help of the feature engineered training and test data.

5.2 DATA CLEANING

At first when we receive the data from the customer in the form of a csv file or a html file, we use the read_csv/read_html function from Pandas library of python to assign the contents of the file as a data frame. If the file is considered to have clean data then we can go to the next step otherwise we must use a heatmap in order to find out the columns where a large amount of null values are present. Box plots are used to find the categorical mean of the of the values which is consequently assigned to the null values based on their categorization. Once data is cleaned, the residual null values and the unwanted columns are removed and the final dataset is attained.

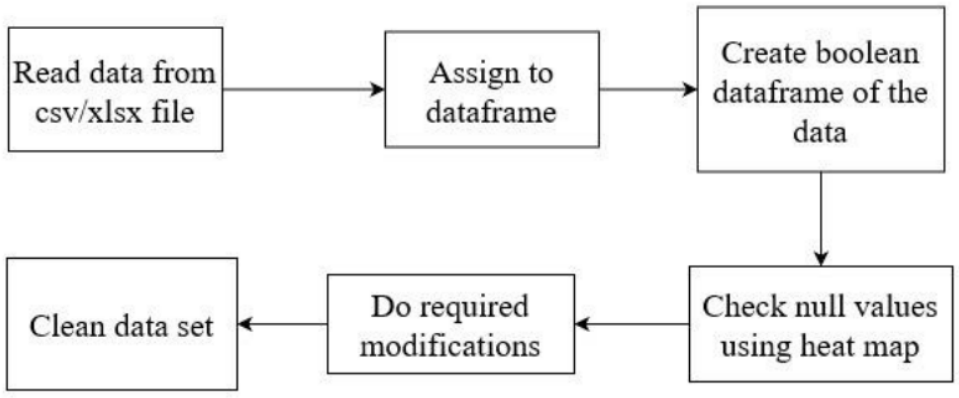


Figure 12:Data Cleaning Block Diagram

For this step we need to follow various sub steps to get our data clean. First, we read the entire data from the csv/xlsx document. Assign the extracted data to the dataframe. Create Boolean values (true or false) of the dataframe, depending on the resultant Boolean value further steps to be followed. Using heatmap function check for all the null values in the dataframe. Extract all the required data and complete all the modifications to be done. Finally, the data is cleaned.

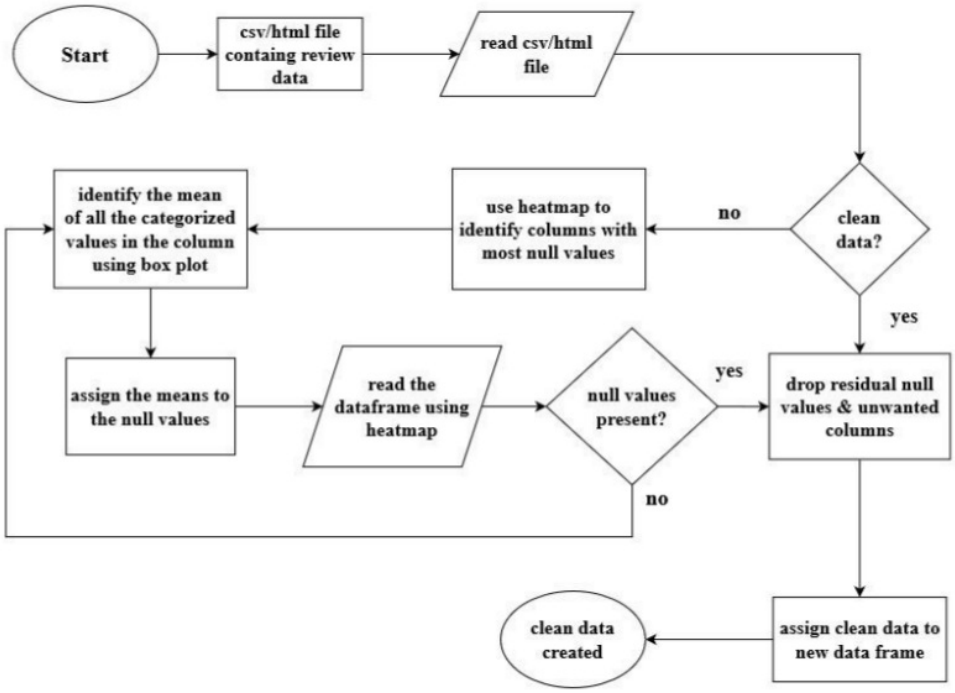
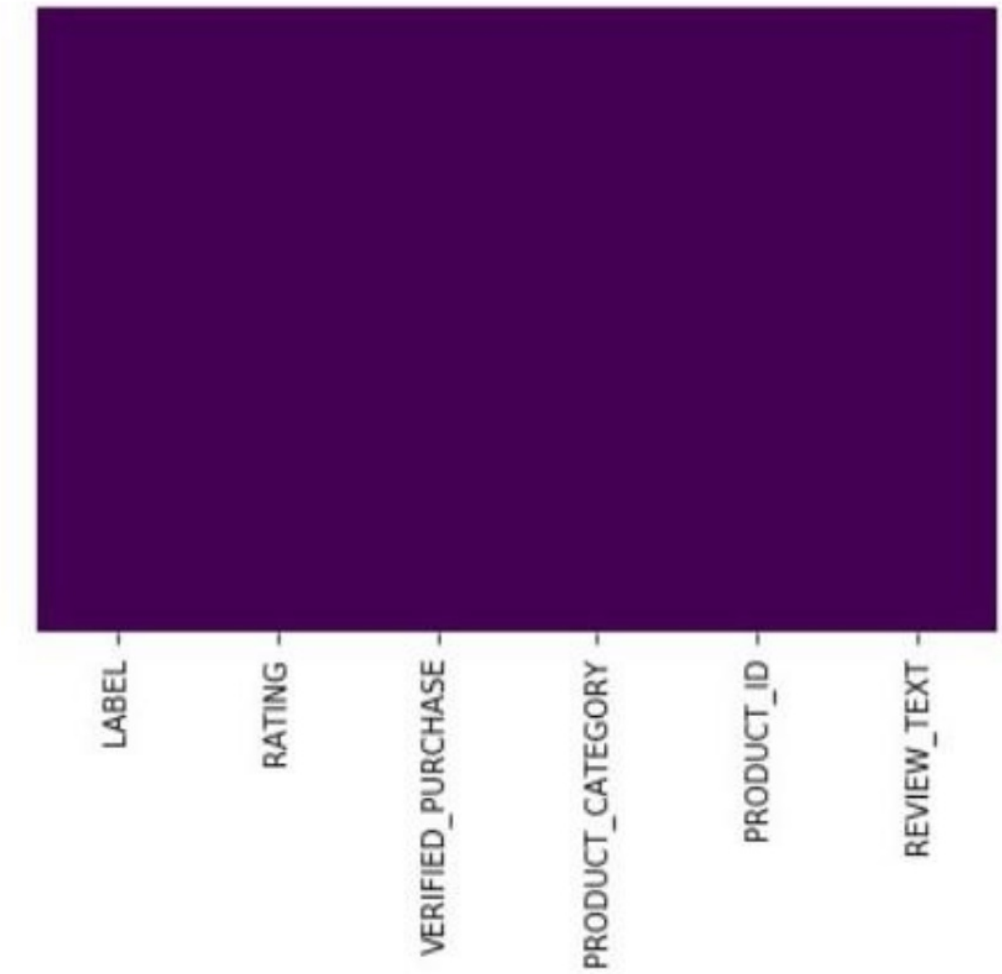


Figure 13:Data Cleaning

	DOC_ID	LABEL	RATING	VERIFIED_PURCHASE	PRODUCT_CATEGORY	PRODUCT_ID	PRODUCT_TITLE	REVIEW_TITLE	REVIEW_TEXT
0	1	Fake	4	N	PC	B00008NG7N	Targus PAUK10U Ultra Mini USB Keypad, Black	useful	When least you think so, this product will sav...
1	2	Fake	4	Y	Wireless	B00LH0Y3NM	Note 3 Battery : Stalion Strength Replacement ...	New era for batteries	Lithium batteries are something new introduced...
2	3	Fake	3	N	Baby	B000I5UZ1Q	Fisher-Price Papasan Cradle Swing, Starlight	doesn't swing very well.	I purchased this swing for my baby. She is 6 m...
3	4	Fake	4	N	Office Products	B003822IRA	Casio MS-80B Standard Function Desktop Calculator	Great computing!	I was looking for an inexpensive desk calcolat...
4	5	Fake	4	N	Beauty	B00PWSAXAM	Shine Whitening - Zero Peroxide Teeth Whitenin...	Only use twice a week	I only use it twice a week and the results are...



	LABEL	RATING	VERIFIED_PURCHASE	PRODUCT_CATEGORY	PRODUCT_ID	REVIEW_TEXT
0	Fake	4	N	PC	B00008NG7N	When least you think so, this product will sav...
1	Fake	4	Y	Wireless	B00LH0Y3NM	Lithium batteries are something new introduced...
2	Fake	3	N	Baby	B000I5UZ1Q	I purchased this swing for my baby. She is 6 m...
3	Fake	4	N	Office Products	B003822IRA	I was looking for an inexpensive desk calcolat...
4	Fake	4	N	Beauty	B00PWSAXAM	I only use it twice a week and the results are...

Figure 14:Data Cleaning Screenshots

5.3 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is method of analyzing datasets. It helps in summing up the overall view of the entire dataset, often with visual representation. This is a technique commonly used by analysts to convey their findings in a clearly understandable format. EDA is commonly done with the help of plots and there are several plots to choose from for unique requirements. Boxplot, Histogram, Scatterplot and Box and Whiskers plot are some of the common types are plots that we use.

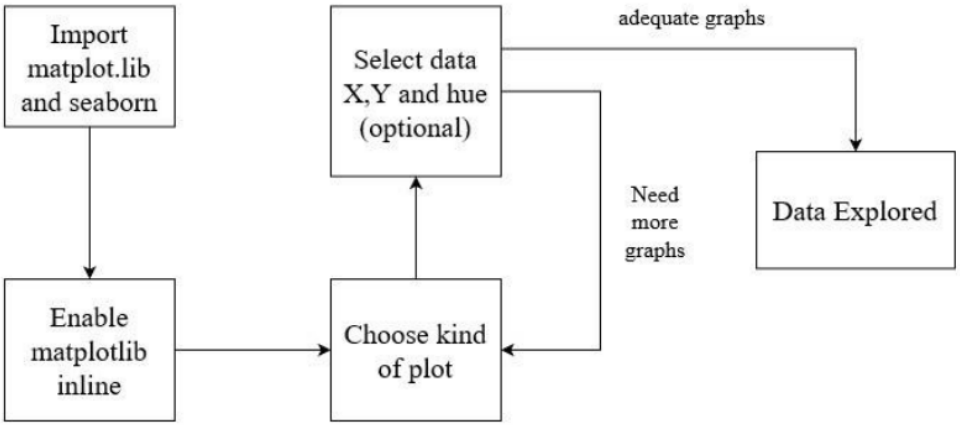


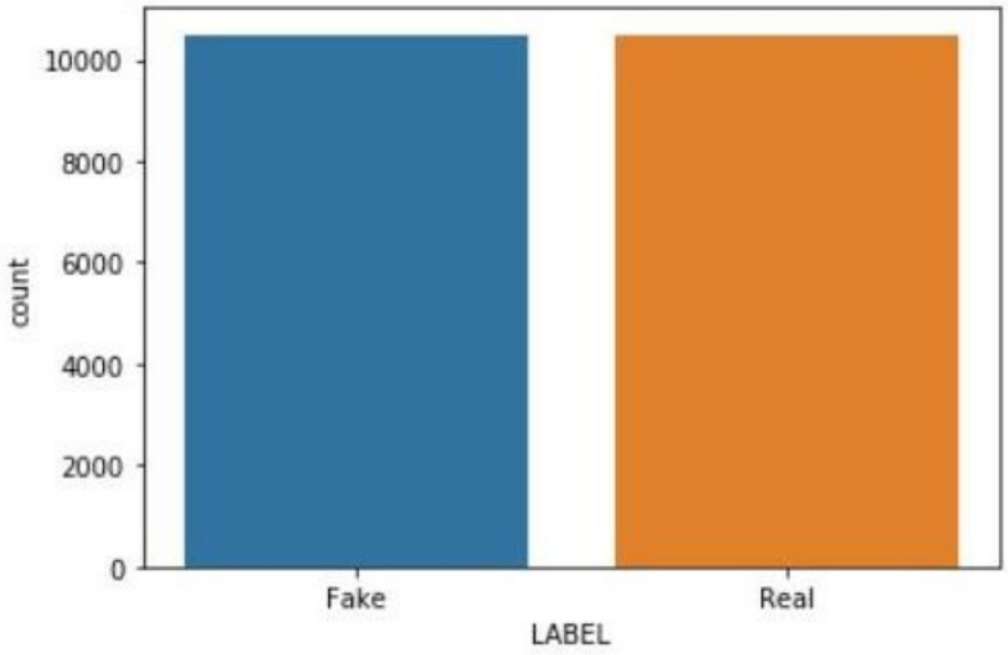
Figure 15:Exploratory Data Analysis

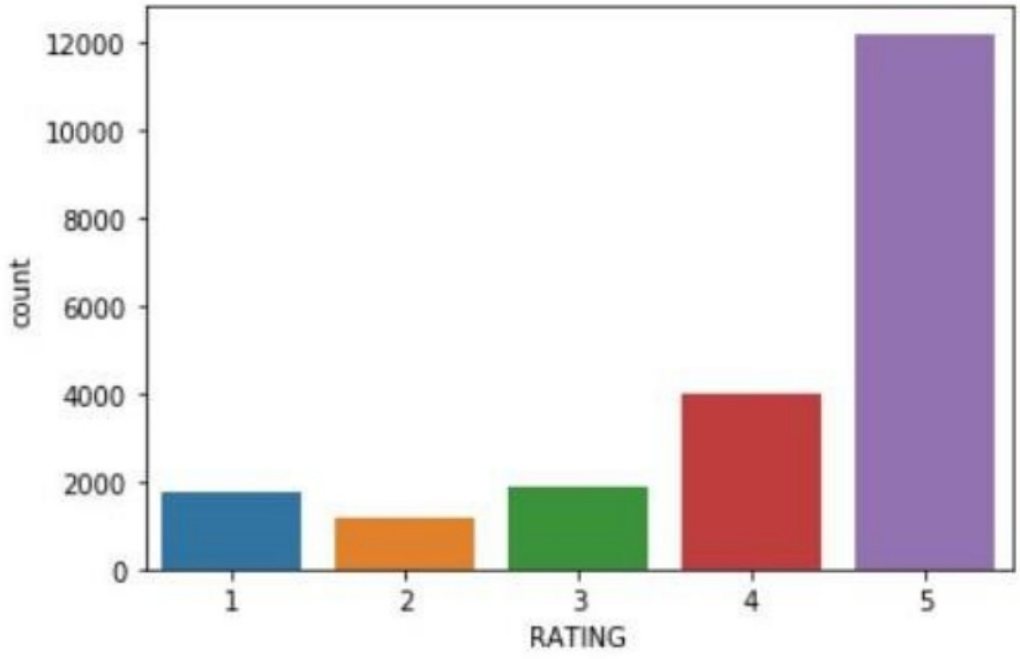
For any program to run successfully we need to get access from some of the header files which have inbuilt functions on which we can perform various operations. We import two files which are

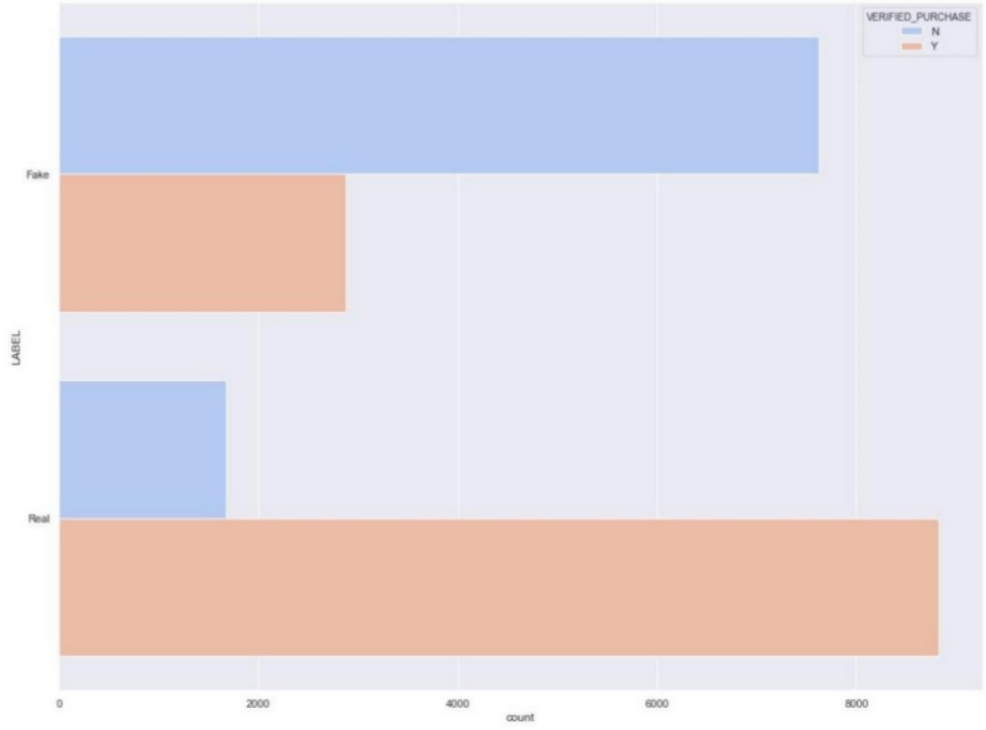
- Matplotlib - a plotting library.
- Seaborn - data visualization library based on matplotlib

We enable matplotlib inline function and choose the type of plot required for our module. We select data X, data Y and hue as an optional plot. Make the graphs and look for satisfactory plots and the data will be explored. If more graphs are required then choose the plot again and make adequate graphs.


12000
10000
8000
6000
4000
2000
0
count
N
Y
VERIFIED_PURCHASE


10000
8000
6000
4000
2000
0
count
Fake
Real
LABEL


12000
10000
8000
6000
4000
2000
0
count
1
2
3
4
5
RATING


VERIFIED_PURCHASE
N
Y
Fake
Real
LABEL
0
2000
4000
6000
8000
count

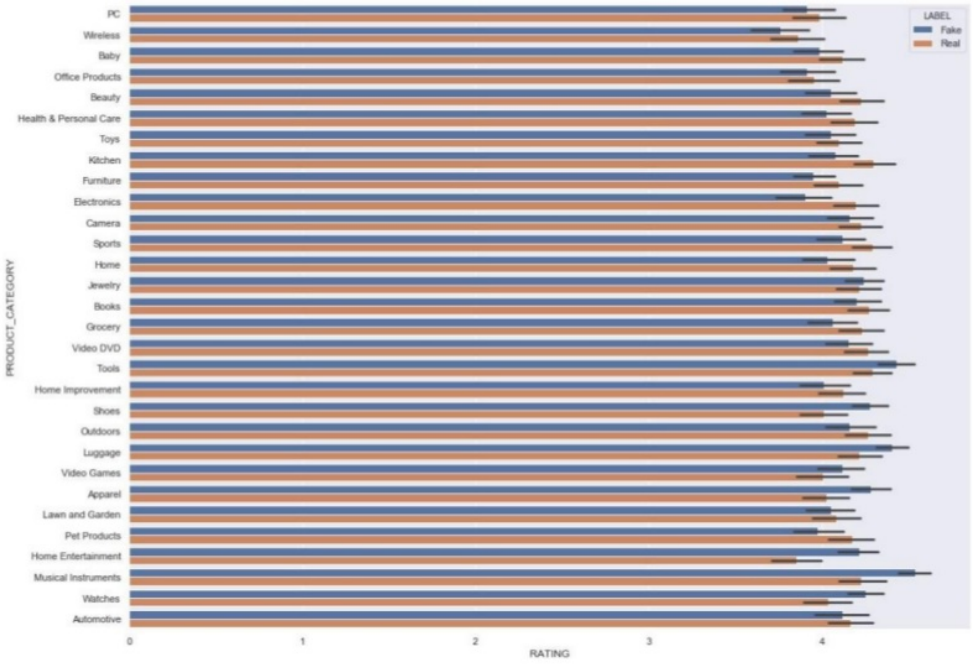


Figure 16:Exploratory Data Analysis Screenshots

5.4 CORPUS

Corpus in general is considered as a sentence whose punctuations, stop words, prefixes as well as suffixes are removed. This method is usually done when dealing with Natural Language Processing which uses the bag of words model. The main purpose of making a Corpus is the make the machine learning model more efficient courtesy of the removal of unwanted words and characters.

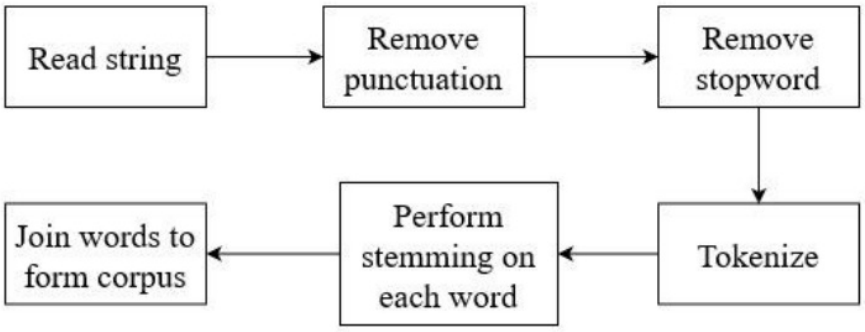


Figure 17:Corpus

First, we read a string as an input in the corpus. We remove all the punctuations present in the string. We remove all the stopwords (words which are not necessary and don't make any difference with or without). Tokenization - very important step in this where we replace the words with a unique identification such as a number to words. We perform stemming of all words mapping them to their root. We join all the leftover words which are essential for the corpus sentence.

```
a=widgets.Combobox(
    value='Shoes',
    placeholder='Choose a category',
    options=cats,
    description='Combobox:',
    ensure_option=True,
    disabled=False
)
display(a)
```

Figure 18:ComboBox

5.4.1 TOKENIZATION

The first basic step in preprocessing data before feeding it into our system is Tokenization. This step is followed in every text classification engine as it helps in reducing the processing time when we train the data. Tokenization involves breaking down of character into individual words called tokens and removing punctuations.

Example for Tokenization:

Input Text: 'A group of top ranked anime villains combined is called Espada.'

Output test: ['A','group','of','top','ranked','anime','villians','combined','is','called','Espada']

This process can now be easily performed in Python by using the .split() function. The next step of Tokenization is Stop-Word Elimination.

5.4.2 STOP-WORD ELIMINATION

Once the characters are broken down into tokens, the next step is Stop-Word Elimination. In this step, all the negative stop words including pronouns, conjunctions and prepositions in the sentence are removed. This is done to reduce the number of words in the text fed into the system without affecting the overall meaning of the sentence, thus in turn reducing the time required for processing when training the data.

Example for Stop-Word Elimination:

Input Text: ['A','group','of','top','ranked','anime','villians','combined','is','called','Espada']

Output Text: ['group','top','ranked','anime','villains','combined','Espada']

Stop-words are removed with the help of 'stopwords' package from nltk.corpus in Python. This package consists of the set of stopwords which if present in the sentence are removed.

5.4.3 STEMMING

Stemming is the process of reducing a word to its root by removing ant prefixes and suffixes in the word. For instance, let us take three words ‘study’, ‘studying’ and ‘studied’, we humans know that these three words have the same meaning but that is not the case with machine. For them these three words have different meaning, so in order to prevent this stemming is done. Stemming converts these three words into ‘studi’ so that the computer can take them as one word

```
0      odd slippers fit great day felt large either s...
1      color bag absolutely amazing inside baglining ...
2      cute buy since earphones always getting tied I...
3      receive many compliments glasses small face lo...
4      hubby pair reef sandals years extremely durabl...
                              ...
695    bought work high arches use arch support heels...
696    Crocs one two brands shoes feet day work Love ...
697    love moccasins fit like custom made mebr soft ...
698    wish little durable got caught bolt crossing b...
699    Ive looking replacement beloved KSO treks owne...
Name: CORPUS, Length: 700, dtype: object
```

	index	LABEL	RATING	VERIFIED_PURCHASE	PRODUCT_ID	REVIEW_TEXT	CORPUS
0	88	Fake	4	Y	B004LBJN84	These are odd slippers, they fit great for a d...	odd slippers fit great day felt large either s...
1	126	Fake	4	N	B00FEW5XZ6	The color of the bag is absolutely amazing! Th...	color bag absolutely amazing inside baglining ...
2	139	Fake	4	Y	B00IHK9EJDS	This is so cute! I have to buy this since my e...	cute buy since earphones always getting tied I...
3	636	Fake	3	N	B005GJ4HH2	I receive so many compliments on these glasses...	receive many compliments glasses small face lo...
4	707	Fake	4	N	B000KJ4FOI	My hubby has had his pair of reef sandals for ...	hubby pair reef sandals years extremely durabl...
...	...	...	...	...	...	...	...
695	20995	Real	4	Y	B00BXYM8T8	I bought these for work. I have high arches, ...	bought work high arches use arch support heels...
696	20996	Real	4	Y	B0014C2ORK	Crocs are one of only two brands of shoes that...	Crocs one two brands shoes feet day work Love ...
697	20997	Real	5	Y	B000EX8CCQ	I love moccasins This fit like it was custom ...	love moccasins fit like custom made mebr soft ...
698	20998	Real	5	Y	B00748YHVE	I wish these were a little more durable. I got...	wish little durable got caught bolt crossing b...
699	20999	Real	4	Y	B00A46KTLU	I've been looking for a replacement for my bel...	Ive looking replacement beloved KSO treks owne...

700 rows × 7 columns

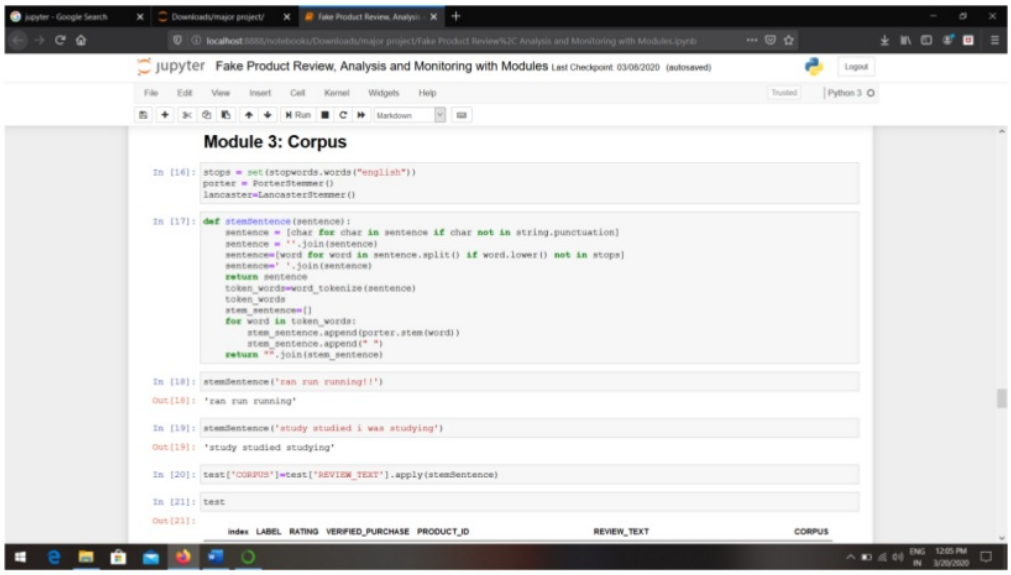


Figure 19: Corpus Screenshots

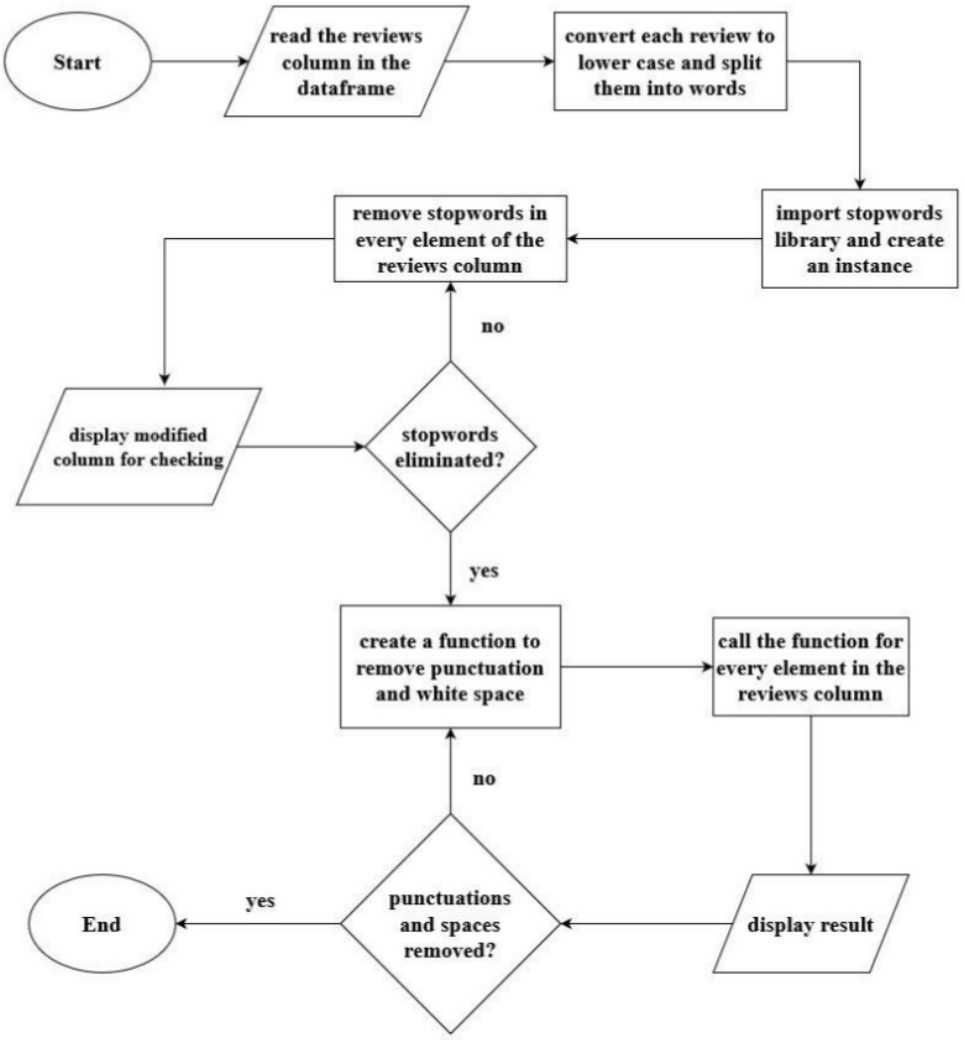


Figure 20:Tokenization and Stop-word elimination

5.5 FEATURE ENGINEERING

Feature engineering is the extraction of data and converting them in a format where the machine learning model can understand. In Layman terms, it is the conversion of all the string values into numbers.

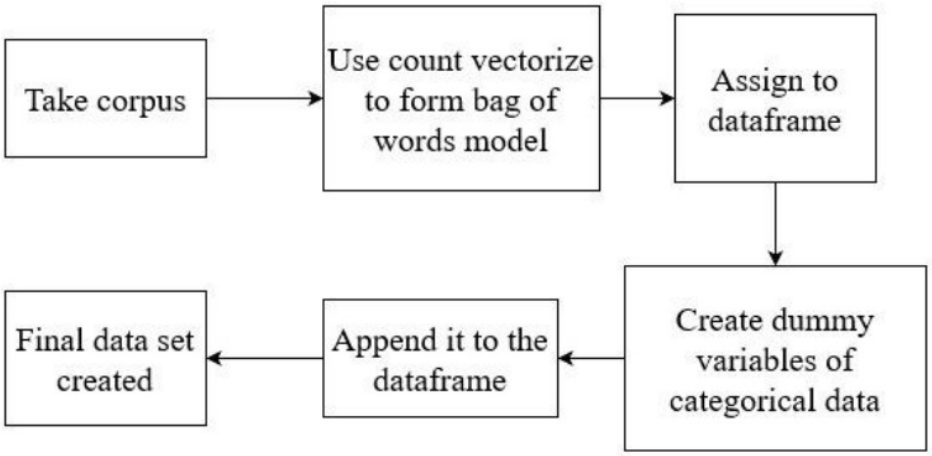


Figure 21:Feature Engineering

In this we carry forward with the corpus sentence. We use the count vectorize function to form a bag of words model. In this model, we extract words which are essential and give them the count depending on the number of times they have appeared. We assign these words to the dataframe. Following this, we create dummy variables of the categorical data where we make categories of different words and append them to the dataframe. Now the final dataset has been created and ready for implementation.

```
bow=vectorizer.fit_transform(corpus)
print(bow.toarray())
print(vectorizer.get_feature_names())
```

```
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
['10', '100', '1000', '100200', '1012', '105', '10s', '11', '110', '1134', '115', '115s', '11s', '12', '12w', '13',
'13th', '14', '15', '16', '17', '180', '1850', '20', '200', '2005', '2014', '23', '247', '25', '2e', '2nd', '30', '
3334', '34', '34air', '34classy34', '34creating', '34dress', '34extra', '34fits', '34heavier34', '34i', '34kit34',
'34large34', '34look34', '34m34', '34made', '34may', '34nice', '34normal34', '34oh', '34overperscription34', '34ove
rsized34', '34pricey34', '34purses34', '34regular34', '34slides34br', '34they', '34we', '34where', '34why', '34wide
', '38', '3899', '38wide', '39', '395', '3br', '3e', '3rd', '3s', '3star', '40', '445', '45', '47', '4wd', '50', '5
00', '5050', '510', '511', '512', '59', '5br', '5hole', '5pm', '5star', '5th', '60', '65', '6900', '734', '75', '75
8', '75ft', '78', '7us40eur', '80', '80s', '810', '83', '85', '850', '856', '885', '8b', '8w', '910', '912', '95',
'995', 'a5', 'ability', 'able', 'abnormal', 'abnormally', 'absolutely', 'absorb', 'absurdly', 'accent', 'accents',
'acceptable', 'accessory', 'accommodate', 'according', 'accurate', 'accurately', 'achilles', 'achy', 'acorn', 'acqu
ired', 'across', 'active', 'activities', 'activity', 'actual', 'actually', 'adaptable', 'adaptor', 'add', 'added',
'adding', 'addition', 'additional', 'additionally', 'adds', 'addso', 'adequate', 'adequately', 'adidas', 'adjust',
```

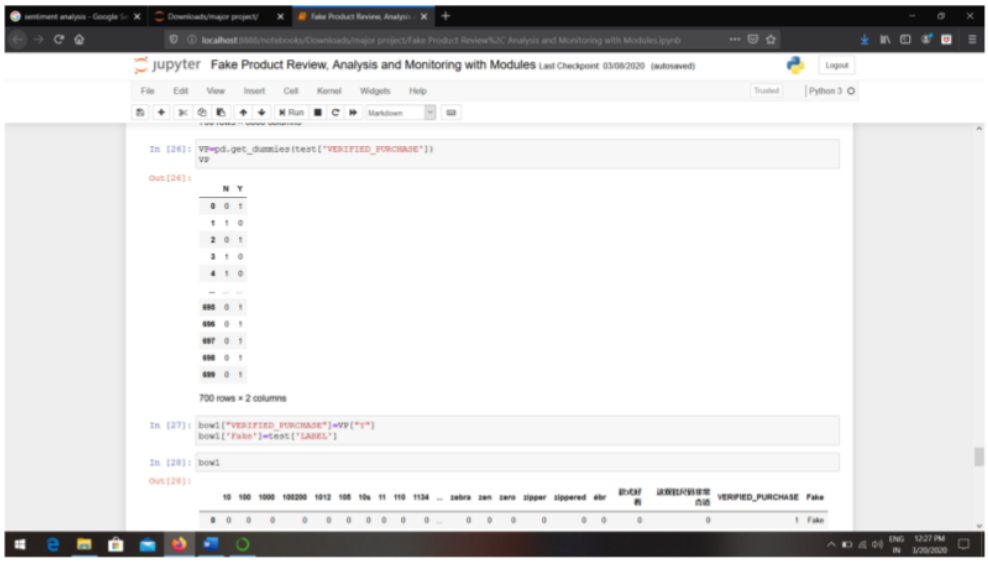


Figure 22:Feature Engineering Screenshots

5.5.1 BAG OF WORDS MODEL

Now that the stop-words are removed, the next step is called the Bag-of-Words Model, or the BoW Model. Until now, beginning from the text being very messy, tokenization was done to break them down and then the negative stop-words were then removed. Now, the BoW model involves calculating the sentence wise frequency of the words present in the whole dataframe. This is like giving a subjectivity score to the text data. That is, a word is a feature in the ML standpoint.

Bag-of-Words is implemented using the concept of sentence_vectors with the help of the nltk package in Python.

Example for Bag-of-Words Model:

['group' = 1,'top' = 1,'ranked' = 0,'anime' = 1,'villians' = 1,'combined' = 0,'Espada' = 1]

"group anime combined" = [1, 1, 0, 1, 0, 1, 1]

"top ranked combined" = [1, 0, 0, 1, 1, 1, 1]

Since the output data is in the form of Boolean values, the processing time and complexity during training of data is very much lower when in comparison to when the output data is in the form of text or string.

5.5.2 DUMMY VARIABLES

Dummy variables are used to convert categorical data into a Numerical Dataframe. This dataframe is then added to the Bag of Words Model Dataframe to form the Final Dataset.

	10	100	1000	100200	1012	105	10s	11	110	1134	...	zebra	zen	zero	zipper	zippered	ébr	款式好看	这双鞋尺码非常合适	VERIFIED_PURCHASE	Fake
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	Fake
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Fake
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	Fake
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Fake
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Fake
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
695	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	Real
696	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	Real
697	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	Real
698	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	Real
699	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	Real

700 rows × 3371 columns

Figure 23:Final Dataset

5.6 NATURAL LANGUAGE PROCESSING

NLP is a technology used for computers to make them understand the normal human language. Humans do not understand binary language and computers do no understand human language so, this processing acts as an intermediate for different analysis.

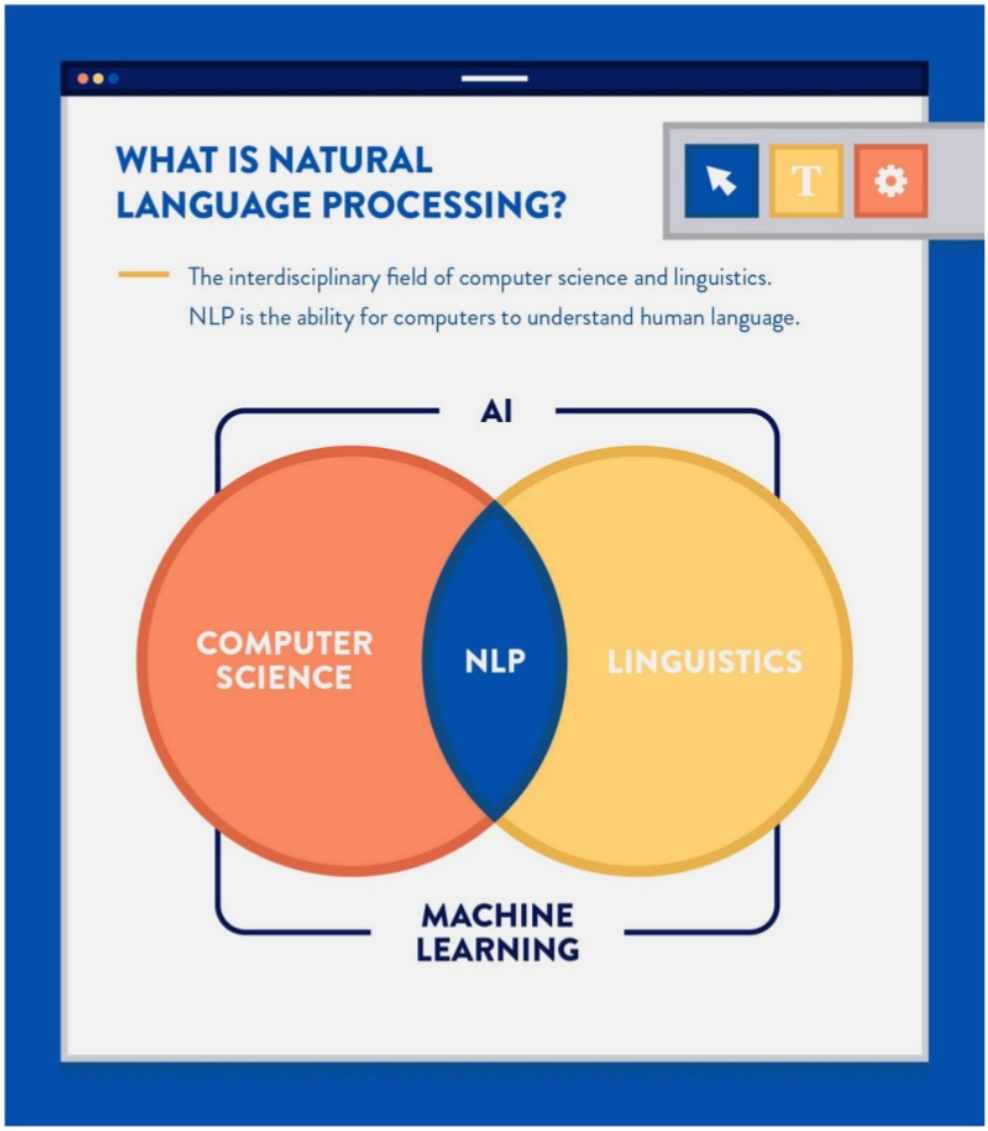


Figure 24:Natural Language Processing

The main objective of NLP is to make the human language to be understood by a computer by ciphered and deciphered text. In fact, a typical interaction between humans and machines using NLP is as follows:

1. A human speaks to a computer
2. The computer stores the human voice as audio
3. Audio is converted to text format
4. NLP for the latest text data

5. Data is converted back to audio format

6. The computer responds back to human in same language

Two major techniques used for completing NLP tasks are syntactic and semantic analysis.

1. Syntactic analysis

For a sentence to be syntactically correct it should be grammatically correct as well. This analysis is used to identify different type of words and sentences are grammatically error free or not. Many types of computer algorithms are used for this type of analysis. Some syntax techniques that can be used:

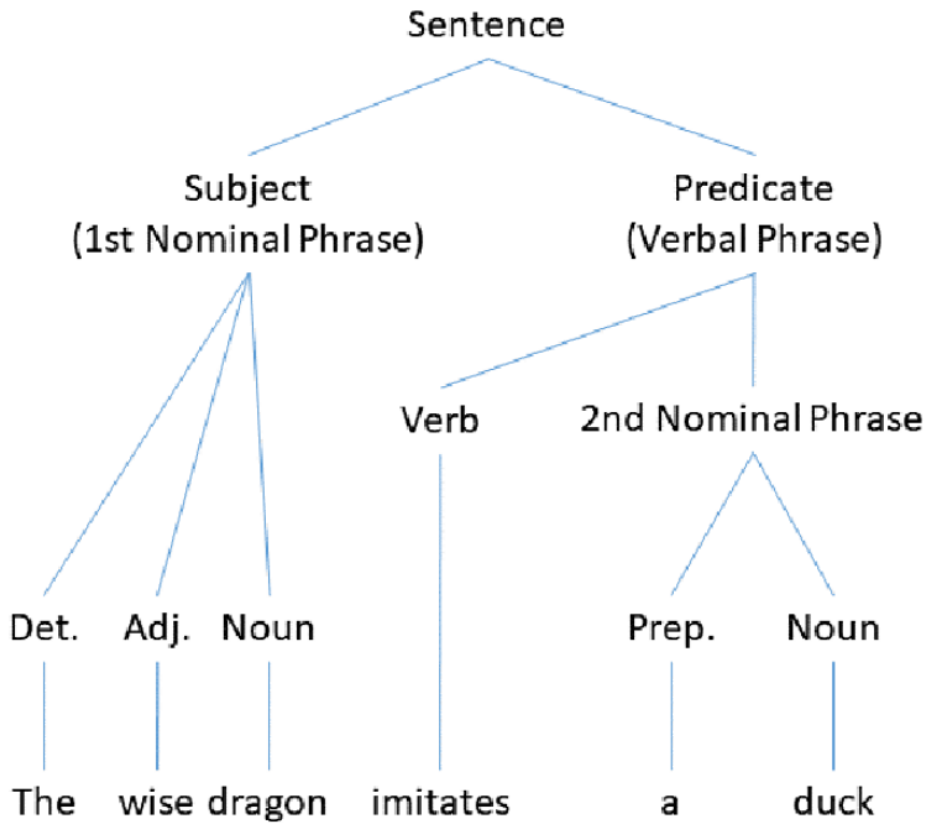


Figure 25:Syntactic Analysis

- **Lemmatization**: Words contain different forms as many words containing multiple meanings. In these, similar forms of words are compiled into a single unit for simple analysis.
- **Morphological segmentation**: They have their unit called morphemes where multiple words are converted into simpler units.
- **Word segmentation**: Large continuous texts or sentences are broken down into small bits and transferred to distinct units.
- **Part-of-speech tagging**: Every word has its own form of speech which is used under this technique.
- **Parsing**: Grammatical check-ups are checked under this technique.
- **Sentence breaking**: Large continuous sentences are made into smaller ones by breaking into multiple sentences.

- **Stemming**: Single form of words are separated into specific distinct units.

2. Semantic analysis

Semantics can be defined as the meaning which is inferred from a text. This analysis is not a completed one, it should be resolved further more for structuring different sentences and words. It is one of the most difficult NLP computer algorithms. Here are some techniques in semantic analysis:

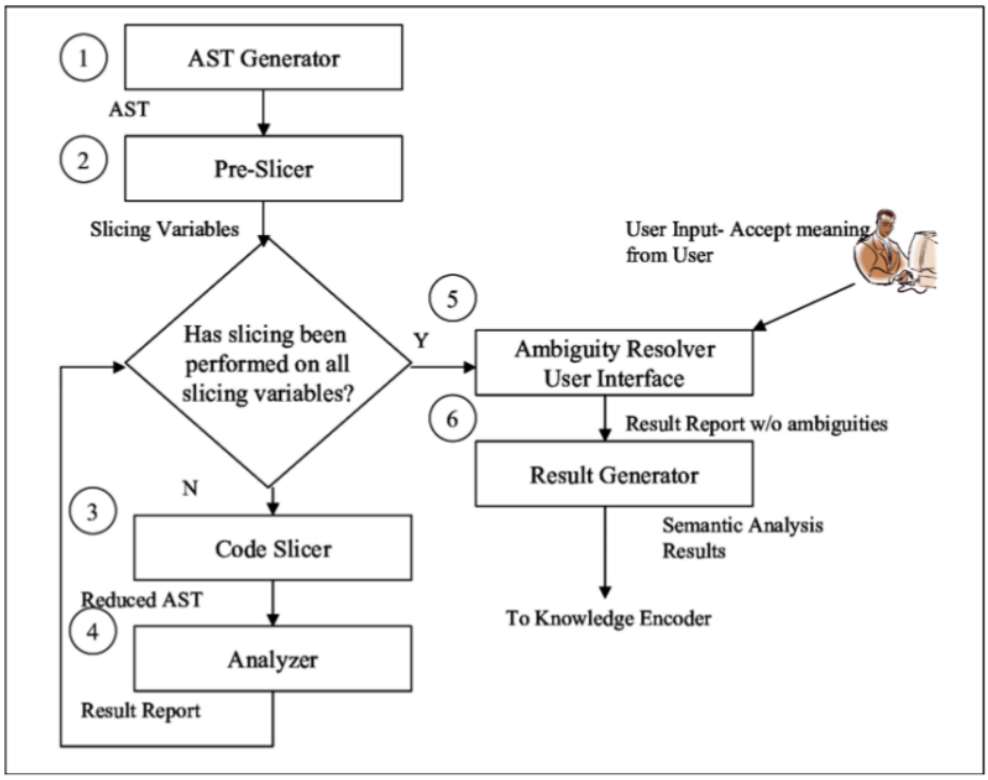


Figure 26:Semantic Analysis

- **Named entity recognition (NER):** It identifies the parts of a sentence which can be adjusted before its use and set them as a group. Such words are names of place, entity, etc.
- **Word sense disambiguation:** It provides meaning to a word which makes sense according to the context.
- **Natural language generation**: It uses databases to determine different semantic intentions from a word and convert them into human language.

5.7 TRAINING THE CLASSIFIER

(1) Training the classifier is a process which involves making our model to learn. That is, it should be trained to give a particular output if given an input. This training process is done

with the help of test samples. A feature vector is obtained from a text using the feature extractor. The tags and feature vectors are combined as a link and fed into an algorithm and generates a new model.

1. Feature Extraction from Text

Text transformation is the first step if we consider any type of machine learning model used to classify text. That along with the bag-of-words model has frequently been used.

Most recently, the new feature vector extraction methods are used dependent on word vectors. These kind of representation of words makes it simpler for various analysis and machine learning algorithms.

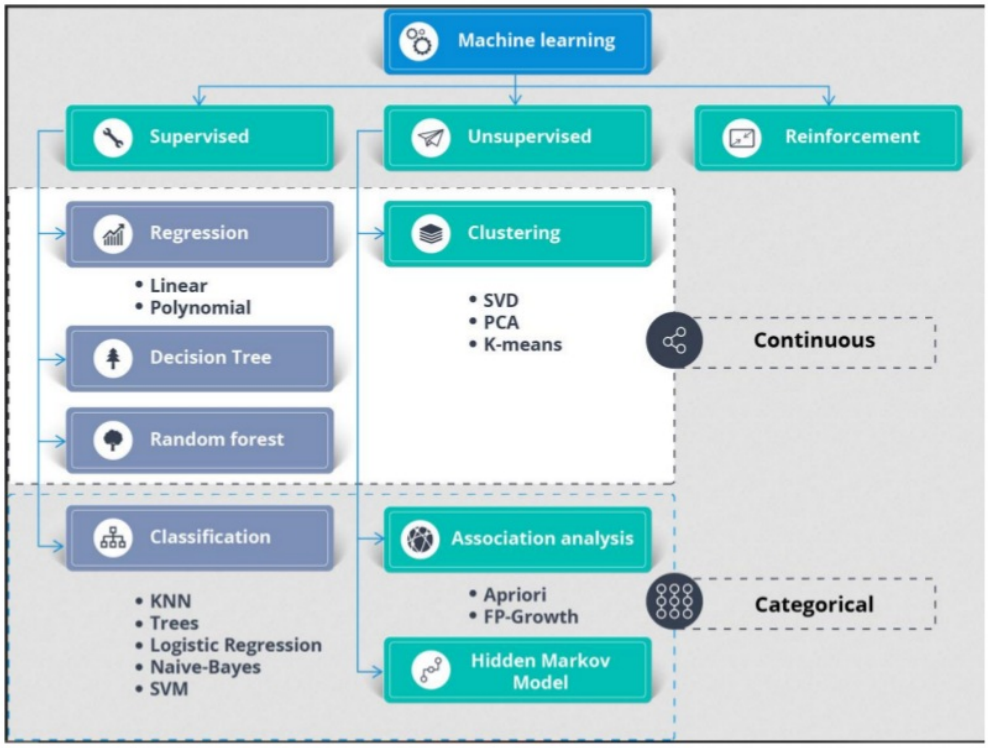


Figure 27:Classification Algorithms

2. Classification Algorithms

They different types of methods are as follows:

- Naïve Bayes: This technique is used for categorizing the text. It is done on the basis various probabilities present.
- Linear Regression: In this technique, the resultant final value is derived based on value of X that is already known value.
- Support Vector Machines: In this technique, the text data is represented in different dimensions. That is, in the second, third and fourth dimension. The Examples are connected to distinct and different regions. The newest data is placed to already existing texts data and to texts they are mapped with.
- Deep Learning: This is a field of neural networks with many diversified algorithms consisting different techniques into it.

5.8 RANDOM FOREST CLASSIFIER

Random Forest Classifier is an ensemble of several independent decision tress that work together to make the predictions required. This is an upgraded version of the normal decision tree model. Each decision tree gives a prediction and the prediction with the highest number will become our final prediction value. Thus, making the required prediction. This is the most efficient way to do binary classification.

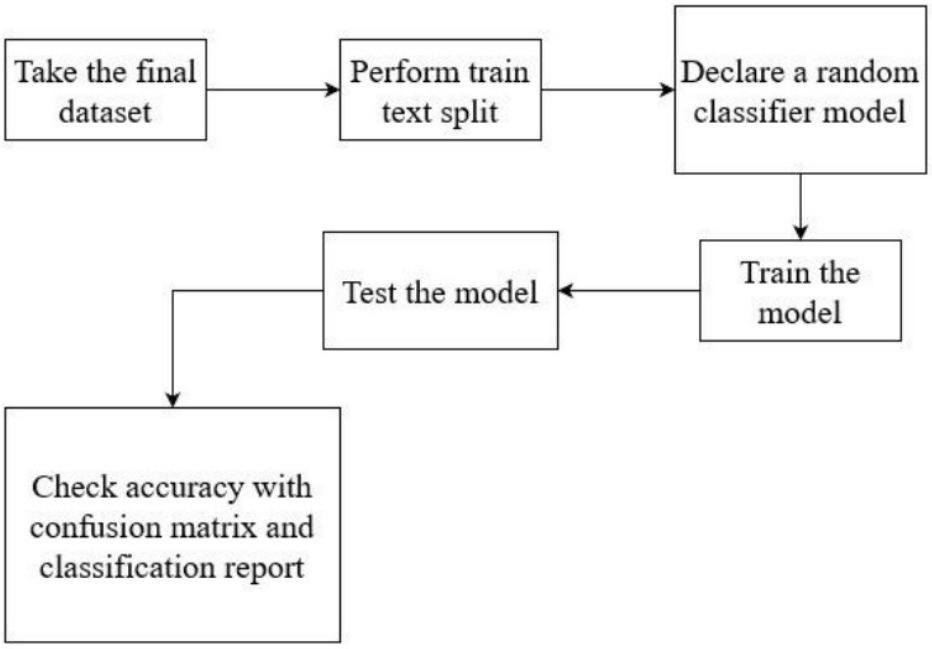
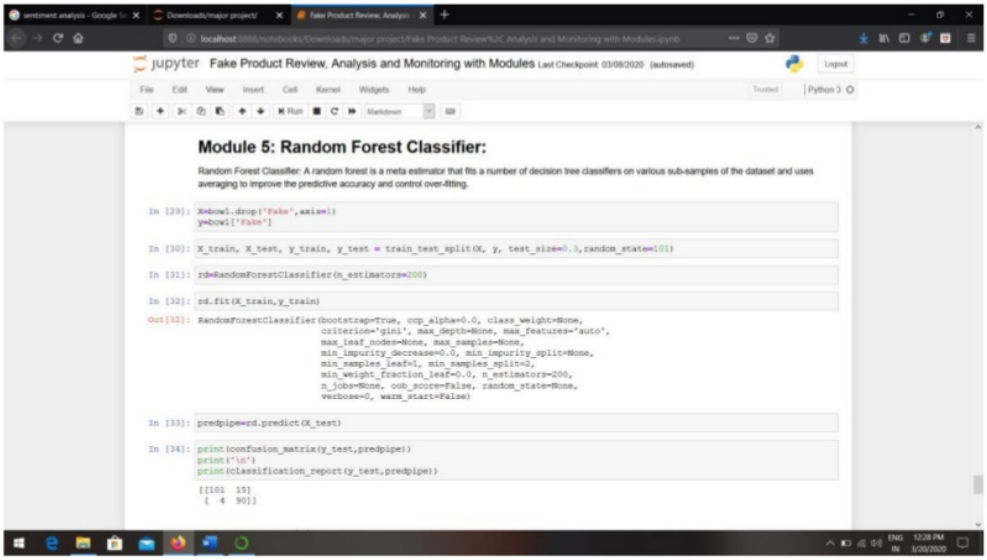


Figure 28:Random Forest Classifier

Having the final dataset, using the train text split function we split the words from the sentences and declare a random classifier model. We train the model depending on our specifications and then test the final model. We check accuracy which defines our model using a confusion matrix and classification report.



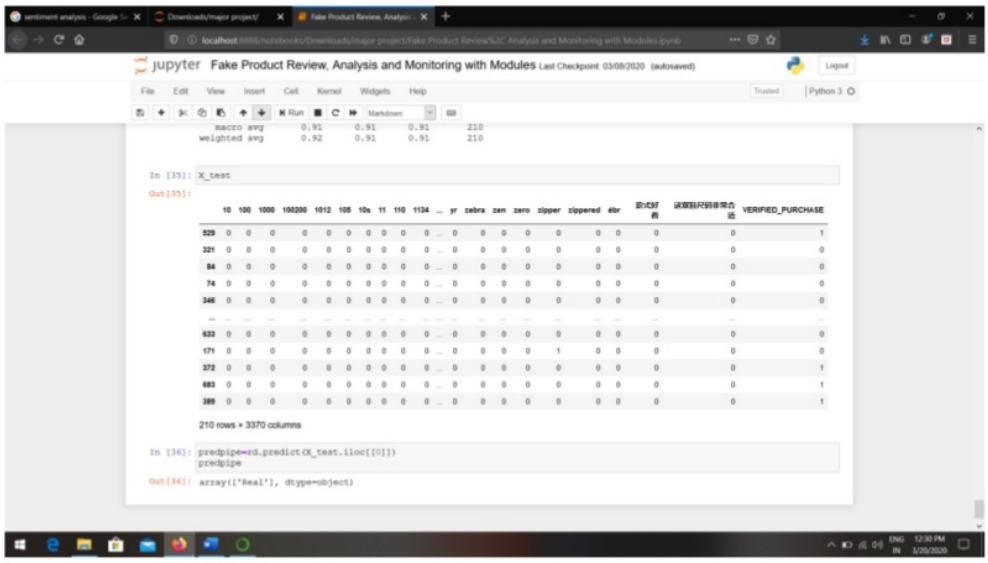


Figure 29:Random Forest Classifier Screenshots

CHAPTER 6

SOCIAL ISSUES AND RESPONSIBILTIES

- Small, unknown and not properly developed products get a five-star rating on e-commerce websites which raises the bar for these products.
- Only these products are shown to customers as their choice of preference depending on the star rating and then customers face problems in the product they order.
- Most of the reviews which give these best ratings to worst products are not verified users. There are huge number of unverified users which are only made to give fake product reviews.
- The quality is not recognized online, customers order them depending on reviews and cheaper rates compared to other websites or offline market.
- If we go through few statistics, survey reports that highest percentage of consumers depend on the reviews and the purchase of the product depends on the reviews given by different consumers.
- Fake reviews are a real-life experience as well. Many customers buy products depending on public recommendations online as well as offline. Whether buying a product or recommending school for students or recommending movie for a moviegoer, everything depends on the reviews given by different reviewers. Some reviews are received successfully which helps to buy a good product but some becomes failure and result in buying a bad product.

CHAPTER 7

DISTRIBUTION OF WORK

Table 1: Distribution of Work

REG NO	NAME	ROLE
RA1611008010005	HARISRIGUHAN S	DATA CLEANING, FEATURE ENGINEERING
RA1611008010029	BHAVESH JAIN N	CORPUS
RA1611008010107	VISHNU KUMAR V H	RANDOM FOREST CLASSIFIER
RA1611008010127	SASIDHARAN R	EXPLORATORY DATA ANALYSIS

CHAPTER 8

WEEK-WISE TIMELINE CHART

Table 2: Week-Wise Timeline Chart

WEEK-WISE DATES	SCHEDULE	STATUS
JAN 5TH To 11TH	Data Preprocessing including cleaning of data. Removal of null values.	COMPLETED.
JAN 12TH To 18TH	Implementation of Tokenization and Stop-Word Elimination. Completion of Heat Maps.	COMPLETED.
JAN 19TH To 25TH	Implementation of Bag-of-Words Model. Reading up on implementing Scrappy.	COMPLETED.
JAN 26TH To 1ST	Completion of Bag-of-Words Model. Documentation for Review start.	COMPLETED.
FEB 2ND TO 8TH	Documentation Work for Review. Diagrams, Tables completion for the Review.	COMPLETED.
FEB 9TH TO 15TH	Training of the data preprocessed. Testing the results using different metrics.	COMPLETED BEFORE SCHEDULE.
FEB 16TH TO 22ND	Implementation of Stemming in corpus for efficiency	COMPLETED BEFORE SCHEDULE.

FEB 23RD TO 29TH	Implement Feature Engineering by turning String values into numerical values using Dummy variables	COMPLETED BEFORE SCHEDULE.
MAR 1ST TO 7TH	Training of Random Forest Classifier model using training data. Completion on Documentation for Review.	COMPLETED BEFORE SCHEDULE.
MAR 8TH TO 14TH	Testing the Model and checking the accuracy using Classification Report and Confusion Matrix. Start working for Plagiarism check.	COMPLETED BEFORE SCHEDULE.
MAR 15TH TO 21ST	Final Documentation work start.	COMPLETED BEFORE SCHEDULE.
MAR 22ND TO 28TH	Completion of Final Documentation for Project.	COMPLETED BEFORE SCHEDULE.

CHAPTER 9

RESULT

We fit the data in the classifier and tested the accuracy of the classifier using precision score, recall score, f1-score and predpipe score. We tested the classifier for Shoes category and the results obtained for the data fitted or fake and real reviews are as follows. The precision scores are 0.95 and 0.87 respectively. The recall scores are 0.87 and 0.95 respectively. The f1-scores are 0.92 and 0.91 respectively. The support scores are 116 and 94 respectively. Thus, proving the accuracy for the classifier that has been fit.

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=200,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)
```

```
a=widgets.Combobox(
    value='Shoes',
    placeholder='Choose a category',
    options=cats,
    description='Combobox:',
    ensure_option=True,
    disabled=False
)
display(a)
```

Combobox: Shoes

```
k=a.value
k
```

```
'Shoes'
```

```
[[102  14]
 [  4  90]]
```

	precision	recall	f1-score	support
Fake	0.96	0.88	0.92	116
Real	0.87	0.96	0.91	94
accuracy			0.91	210
macro avg	0.91	0.92	0.91	210
weighted avg	0.92	0.91	0.91	210

```
a=widgets.Combobox(
    value='Shoes',
    placeholder='Choose a category',
    options=cats,
    description='Combobox:',
    ensure_option=True,
    disabled=False
)
display(a)
```


```
k=a.value
k
```

```
'Shoes'
```

```
test = pd.DataFrame(df[df["PRODUCT_CATEGORY"]==k
```

```
a=widgets.Combobox(
    value='Shoes',
    placeholder='Choose a category',
    options=cats,
    description='Combobox:',
    ensure_option=True,
    disabled=False
)
display(a)
```

Combobox: PC

```
k=a.value
k
```

```
'PC'
```

```
[[76 40]
 [13 81]]
```

	precision	recall	f1-score	support
Fake	0.85	0.66	0.74	116
Real	0.67	0.86	0.75	94
accuracy			0.75	210
macro avg	0.76	0.76	0.75	210
weighted avg	0.77	0.75	0.75	210

```
X_test
```

	01022016	03	10	100	1000	1000br	100v	101	10104	1034	...	youve	yr	z77ma	z77mag45	zelotes	zero	zeroed	zipper	item	VERIFIED_PURCHASE
529	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
321	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
84	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
74	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
346	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
633	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
171	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
372	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
683	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
389	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1

210 rows × 5370 columns

```
predpipe=rd.predict(X_test.iloc[[0]])
predpipe
```

```
array(['Real'], dtype=object)
```

```
predpipe=rd.predict(X_test.iloc[[50]])
predpipe
```

```
array(['Fake'], dtype=object)
```

Figure 30:Result Screenshots

CHAPTER 10

CONCLUSION

Through this paper we detect the presence of fake reviews in the online shopping system using Amazon reviews dataset. This is done with the help of Random Forest Classifier. We chose Random Forest Classifier over other methods like Sentiment Analysis and Opinion mining due to the time effectiveness it shows dominating the latter two. Processes of Corpus creation, Feature engineering and Bag of Words model have helped in achieving the time effectiveness needed. The model gives an accuracy ranged from 75-95% depending on the product category that the user wants to choose from. The accuracy of the detection of fake reviews are solely based on the product type itself and their variants. This will allow users to choose from the particular category that will want their reviews from.

CHAPTER 11

FUTURE ENHANCEMENT

In the future, we have planned to make this project more dynamic. Currently, we use an already existing dataset and our findings and analysis are derived from that particular dataset. In the future, we plan to make it more dynamic by allowing the user to choose the website URL on their own from a product they wish to buy directly from Amazon. They will have to copy and paste the URL, from which the analysis of data will be done and the real and fake reviews will be detected. Thus, providing analysis of real-world data. This can be done with the help of a python Library called scrappy which will parse and scrap webpages on its own if provided the URL, thus making our project more dynamic. Further, we are planning on making our code as an extension that can run on any browser. Here the extension or add-on, when we reach a particular e-commerce site, automatically scans the page and shows the amount of real and fake reviews that are present. Thus, making the process much more trivial for users to make use of.

CHAPTER 12

REFERENCES

[1] Anusha Sinha, Nishant Arora, Shipra Singh, Mohita Cheema, Akthar Nazir," Fake Product Review Monitoring Using Opinion Mining", International Journal of Pure and Applied Mathematics, Volume 119, No. 12, 2018, pp.13203-13209

[2] Maria Soledad Elli, Yi-Fan Wang," Amazon Reviews, business analytics with sentiment analysis"

[3] Amit Sawant, Shraddha Bhange, Shruti Desai, Akash Pandey," Fake Product Review Monitoring and Removal for Proper Ratings", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 8 Issue 03, March-2019

[4] Piyush Jain, Karan Chheda, Mihir Jain, Prachiti Lade, "Fake Product Review Monitoring System", International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3, Issue-3, April 2019, pp. 105-107

[5] Madhura N Hegde, Sanjeetha K Shetty, Sheikh Mohammed Anas, Varun K, "FAKE PRODUCT REVIEW MONITORING", International Research Journal of Engineering and Technology (IRJET) ,e-ISSN: 2395-0056, Volume 05, Issue 06, June 2018

[6] Manleen Kaur Kohli, Shaheen Jamil Khan, Tanvi Mirashi, Suraj Gupta," Fake Product Review Monitoring and Removal for Genuine Online Product Reviews Using Opinion Mining", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 7, Issue 1, January 2017

[7] Xinkai Yang, "One Methodology for Spam Review Detection Based on Review Coherence Metrics", International Conference on Intelligent Computing and Internet of Things (IC1T) 2015.

[8] Xing Fang and Justin Zhan," Sentiment analysis using product review data", Fang and Zhan Journal of Big Data (2015).

[9] NJindal and B.Liu, "Review spam detection", In Proceedings of the 16th International Conference on World Wide Web, pp1189-1190, 2007

[10] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Review graph based online store review spammer detection", IEEE 11th International Conference on Data Mining, pp 1242-1247,2011.

[11] Liu, Bing. Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies 5.1 (2012): 1-167

[12] Review ranking method for spam recognition, Jamia Millia Islamia New Delhi India, gunjan_ansari@yahoo.com, tahmad2@jmi.ac.in, ndoja@yahoo.com

[13] Rajashree S. Jadhav, Prof. Deipali V. Gore, "A New Approach for Identifying Manipulated Online Reviews using Decision Tree ". (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), pp 1447-1450, 2014

[14] Long- Sheng Chen, Jui-Yu Lin, "A study on Review Manipulation Classification using Decision Tree", Kuala Lumpur, Malaysia, pp 3-5, IEEE conference publication, 2013

[15] SP.Rajamohana, Dr.K.Umamaheshwari, M.Dharani, R.Vedackshya, "A survey on online review SPAM detection techniques", International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT) 2017, ISBN(e): 978-1-5090-5778-8. ", International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT) 2017, ISBN(e): 978-1-5090-5778-8.

[16] Cambria, E; Schuller, B; Xia, Y; Havasi, C (2013). "New avenues in opinion mining and sentiment analysis". IEEE Intelligent Systems. 28 (2): 15 21.doi:10.1109/MIS.2013.30.

[17] Michael Beaney (Summer 2012). "Analysis". *The Stanford Encyclopedia of Philosophy*. Michael Beaney. Retrieved 23 May 2012.

[18] Jeneen Interlandi (February 8, 2010). "The fake-food detectives". *Newsweek*. Archived from the original on October 21, 2010.

[19] Benjamin Snyder and Regina Brazil, “Multiple Aspect ranking using the Good Grief Algorithm “Computer Science and Artificial Intelligence Laboratory Massachusetts Institute of Technology2007.

[20] Ivan Tetovo, “A Joint Model of Text and Aspect Ratings for Sentiment Summarization “Ivan Department of Computer Science University of Illinois at Urbana, 2011

[21] McAuley, Julian, et al. Image-based recommendations on styles and substitutes Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2015.

[22] Hovy, Dirk, Anders Johannsen, and Anders Sgaard. User review sites as a resource for large-scale sociolinguistic studies., Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015.

[23] Dickinson, Brian, and Wei Hu. Sentiment Analysis of Investor Opinions on Twitter., Social Networking 4.03 (2015): 62.

[24] Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. Modern information retrieval., Vol. 463. New York: ACM press, 1999.

[25]Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schtze. Introduction to information retrieval. Vol. 1., Cambridge: Cambridge university press, 2008.

[26] Joachims, Thorsten Text categorization with support vector machines: Learning with many relevant features., Springer Berlin Heidelberg, 1998.

[27]Wu, Ting-Fan, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling., The Journal of Machine Learning Research 5 (2004): 975-1005.

[28] Dea Delvia Arifin,, Shaufiah and Moch. Arif Bijaksana," Enhancing Spam Detection on Mobile Phone Short Message Service (SMS) Performance using FP-Growth and Naive Bayes Classifier", The 2016 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob).

[29] Han, Jiawei, Jian Pei, and Yiwen Yin. "Miningfrequent patterns without candidate generation." ACM Sigmod Record. Vol. 29. No. 2. ACM, 2000.

[30] Gann W-JK, Day J, Zhou S (2014) Twitter analytics for insider trading fraud detection system. In: Proceedings of the sencond ASE international conference on Big Data. ASE

[31] Roth D, Zelenko D (1998) Part of speech tagging using a network of linear separators. In:Coling-Acl, The 17th International Conference on Computational Linguistics.

[32] Zhang Y, Xiang X, Yin C, Shang L (2013) Parallel sentiment polarity classification method with substring feature reduction. In: Trends and Applications in Knowledge Discovery and Data Mining, volume 7867 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, Heidelberg, Germany.

[33] Ranks.nl, "Stopwords". [Online]. Available: http://www.ranks.nl/stopwords.

[34] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques 3rd Edition. Morgan Kaufmann Publishers, 2013.

[35] Tan LK-W, Na J-C, Theng Y-L, Chang K (2011) Sentence-level sentiment polarity classification using a linguistic approach. In: Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation. Springer, Heidelberg, Germany.

[36] Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA.

APPENDIX

```
import numpy as np
import nltk
import string
import bs4 as bs
import re
import pandas as pd
import matplotlib as plt
import seaborn as sns
%matplotlib inline
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer,LancasterStemmer
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import CountVectorizer,TfidfTransformer
from sklearn.ensemble import RandomForestClassifier
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import confusion_matrix,classification_report
import ipywidgets as widgets
from ipywidgets import interact, interact_manual
df=pd.read_excel("amazon_reviews.xlsx")
df.head()
del df['DOC_ID']
del df['PRODUCT_TITLE']
del df['REVIEW_TITLE']
df.head()
sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis')
df.dropna()
df
sns.countplot(df['VERIFIED_PURCHASE'])
sns.countplot(df['LABEL'])
```

```
sns.countplot(df['RATING'])
sns.set(rc={'figure.figsize':(17,13)})
sns.barplot(x='RATING',y='PRODUCT_CATEGORY',data=df,hue='LABEL')
sns.countplot(y='LABEL',data=df,palette='coolwarm',hue='VERIFIED_PURCHASE')
cats=list(df['PRODUCT_CATEGORY'].unique())
cats
a=widgets.Combobox(
    value='Shoes',
    placeholder='Choose a category',
    options=cats,
    description='Combobox:',
    ensure_option=True,
    disabled=False
)
display(a)
k=a.value
k
test =
pd.DataFrame(df[df["PRODUCT_CATEGORY"]==k].drop('PRODUCT_CATEGORY',a
xis=1))
test.columns =
["LABEL","RATING","VERIFIED_PURCHASE","PRODUCT_ID","REVIEW_TEXT"]
test.reset_index(inplace=True)
test
stops = set(stopwords.words("english"))
porter = PorterStemmer()
lancaster=LancasterStemmer()
def stemSentence(sentence):
    sentence = [char for char in sentence if char not in string.punctuation]
    sentence = ''.join(sentence)
    sentence=[word for word in sentence.split() if word.lower() not in stops]
    sentence=' '.join(sentence)
    return sentence
    token_words=word_tokenize(sentence)
```

```
    token_words
    stem_sentence=[]
    for word in token_words:
        stem_sentence.append(porter.stem(word))
        stem_sentence.append(" ")
    return "".join(stem_sentence)
test['CORPUS']=test['REVIEW_TEXT'].apply(stemSentence)
test
corpus = test['CORPUS']
corpus
vectorizer=CountVectorizer()
bow=vectorizer.fit_transform(corpus)
print(bow.toarray())
print(vectorizer.get_feature_names())
bow1=pd.DataFrame(bow.toarray(),columns=vectorizer.get_feature_names())
bow1
VP=pd.get_dummies(test['VERIFIED_PURCHASE'])
VP
bow1["VERIFIED_PURCHASE"]=VP["Y"]
bow1['Fake']=test['LABEL']
bow1
X=bow1.drop('Fake',axis=1)
y=bow1['Fake']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,random_state=101)
rd=RandomForestClassifier(n_estimators=200)
rd.fit(X_train,y_train)
predpipe=rd.predict(X_test)
print(confusion_matrix(y_test,predpipe))
print('\n')
print(classification_report(y_test,predpipe))
X_test
predpipe=rd.predict(X_test.iloc[[0]])
predpipe
```

PAPER PUBLICATION STATUS

Submitted to a conference and accepted.

Major_Project_Report.pdf

ORIGINALITY REPORT

5%	3%	3%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dwhsys.com Internet Source	1%
2	es.scribd.com Internet Source	1%
3	www.datacamp.com Internet Source	<1%
4	K. G. Srinivasa, Siddesh G. M., Srinidhi H.. "Network Data Analytics", Springer Science and Business Media LLC, 2018 Publication	<1%
5	Submitted to Trinity College Dublin Student Paper	<1%
6	Lecture Notes in Computer Science, 2012. Publication	<1%
7	Submitted to University of Computer Studies Student Paper	<1%
8	"Advances in Computer Communication and Computational Sciences", Springer Science and Business Media LLC, 2019	<1%

	Publication	
9	www.sportademics.com Internet Source	<1%
10	Submitted to University of Bristol Student Paper	<1%
11	Submitted to Manchester Metropolitan University Student Paper	<1%
12	Submitted to Glyndwr University Student Paper	<1%
13	Submitted to The University of Manchester Student Paper	<1%
14	etd.wvu.edu Internet Source	<1%
15	studentsrepo.um.edu.my Internet Source	<1%
16	Submitted to University College London Student Paper	<1%
17	augustineventures.com Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		