**SENTIMENT ANALYSIS USING NLP**

**MAJOR PROJECT REPORT**

***Submitted by***

**MOHD SAEED AFRI**

TABLE OF FIGURES

|  |
| --- |
| [Figure 1 : Count Vectorizer Matrix 7](#_Toc104581959)  [Figure 2: TFIDF Formula 7](#_Toc104581960)  [Figure 3: Multinomial Naive Bayes 9](#_Toc104581961)  [Figure 4: Dataset Description 10](#_Toc104581962)  [Figure 5: Methodology Flowchart 12](#_Toc104581963)  [Figure 6: Method 12](#_Toc104581964)  [Figure 7: Pipeline 13](#_Toc104581965)  [Figure 8: Distribution 14](#_Toc104581966)  [Figure 9: Helpfulness Distribution 14](#_Toc104581967)  [Figure 10: Frequency 15](#_Toc104581968)  [Figure 11: Review Score Distributions 15](#_Toc104581969)  [Figure 12: Review distribution 16](#_Toc104581970)  [Figure 13: NLP Steps 18](#_Toc104581971)  [Figure 15: Bayes Rule 19](#_Toc104581973)  [Figure 16:Cross Validation 21](#_Toc104581974)  [Figure 17: confusion matrix 22](#_Toc104581975)  [Figure 18:Confusion Matrix 22](#_Toc104581976)  [Figure 19: Accuracy All Tables 23](#_Toc104581977)  [Figure 20: ROC Curve 23](#_Toc104581978)  [Figure 21: AUC Curve 26](#_Toc104581979)  [Figure 22: AUC 26](#_Toc104581980)  [Figure 23: Table 26](#_Toc104581981)  [Figure 24: AUC KNN TFIDF 27](#_Toc104581982)  [Figure 25: ROC Curve 27](#_Toc104581983)  [Figure 26: Accuracy 27](#_Toc104581984)  [Figure 27: AUC KNN 28](#_Toc104581985)  [Figure 28: ROC KNN 28](#_Toc104581986)  [Figure 29: Final KNN 28](#_Toc104581987)  [Figure 30: AUC KNN KD Tree 29](#_Toc104581988)  [Figure 31: ROC Curve 29](#_Toc104581989)  [Figure 32: Final knn kdtree 29](#_Toc104581990)  [Figure 33: Final Reports 30](#_Toc104581991) |

INDEX

|  |  |  |
| --- | --- | --- |
| S.NO | TOPICS | PAGE NO |
| 1 | DECLARATION | 5 |
| 2 | ABSTRACT | 6 |
| 3 | PROBLEM DEFINITION | 8 |
| 4 | COUNT - VECTOR | 9 |
| 5 | TF-IDF VECTORIZER | 9 |
| 6 | SUMMARY FEATURES | 9 |
| 7 | ALGORITHMS | 10 |
| 8 | DATASET DESCRIPTION | 11 |
| 9 | REVIEW AND SUMMARY TEXT | 12 |
| 10 | METHODOLOGY | 13 |
| 11 | EXPLORATORY DATA ANALYSIS | 14 |
| 12 | TEXT PREPROCESSING | 16 |
| 13 | TOKENIZATION | 19 |
| 14 | CONVERTING UPPERCASE TOKENS INTO LOWERCASE | 19 |
| 15 | REMOVAL OF PUNCTUATION MARKS | 19 |
| 16 | STOP WORDS | 19 |
| 17 | LEMMATIZATION | 19 |
| 18 | MODEL TRAINING | 19 |
| 19 | MODEL VALIDATION | 21 |
| 20 | ANALYSIS OF FEATURES | 22 |
| 21 | REFERENCES | 32 |
| 22 | PLAGAIRISM REPORT | 33 |

## ABSTRACT

Consumers may determine the quality of their purchases by reading and scrutinizing reviews made by users of online shopping sites on social media. Furthermore, internet marketplaces such as Amazon.com allow “helpful" if the content of the review is useful to them.

This allows both consumers and producers to analyze broad preferences more quickly by focusing on a few key judgments. Recent reviews, on the other hand, have a relatively low number of votes, with highly rated reviews appearing first on the user's radar.

This study addresses these problems by developing a text classification which helps to identify its usefulness regardless of when online reviews are published. This survey is based on data collected by Amazon.com, which consists of delicatessen reviews.  
  
The main focus of research so far has been to identify a relation among degree of usefulness of a review and the content-based characteristics of the review.

This study combines three separate ways to predict the value of reviews, including vectorized traits, review-centric traits, and summary-centric features, in addition to uncovering relevant content-based qualities. Increase. In addition, classic text categorization classifiers such as multinomial naive Bayes, k-nearest neighbor method with and without kd tree method. Vectorized features outperform other features by a significant margin.

**INTRODUCTION**

Companies are now selling both products as well as their services through online places and social media. The growing usage of social media, Internet has drastically altered the way people shop for goods. Customers can also get product information through social media.

Customers trust online consumer evaluations more than vendor information, according to studies. Furthermore, online customer evaluations are more user-centric, describing the product from the perspective of the user.  
Consumers may assess the quality of their items by reading and evaluating the various online reviews accessible on public sites like Amazon.com. However, if these reviews are not properly arranged or displayed, buyers may find themselves in an unclear scenario. The overwhelming amount of reviews can affect consumers' impartial decision-making, primarily due to uneven or untrue (spam) review quality.  
To address this problem, Amazon.com includes a feature in the review text that allows readers to rate how beneficial the information supplied in the review is. Consumers may use this as a guide to help them make informed purchasing decisions.

## 

## PROBLEM DEFINITION

This project's purpose is to create an automatic text-based categorization system that can reliably forecast the utility of Amazon's online consumer reviews. The task at hand is to conduct binary classification utilizing an interconnection of text descriptions and ML methods.

"1" denotes "helpful" and "0" denotes "useless" in the binary class. The amount of people who evaluate the review as "helpful" determines the usefulness scale.

The prediction model is built using publicly accessible training data, which includes product review texts, summaries, ratings, usefulness rating details, and other product and user-related data.

The classification algorithms are as follows:

* k-Nearest Neighbors
* Naive Bayes.

This project also describes how to preprocess the required text.

## VECTORIZED FEATURES

A vectorization technique is used to turn a bundle of textual data to numeric feature vector in this group of features.

### COUNT – VECTORIZER

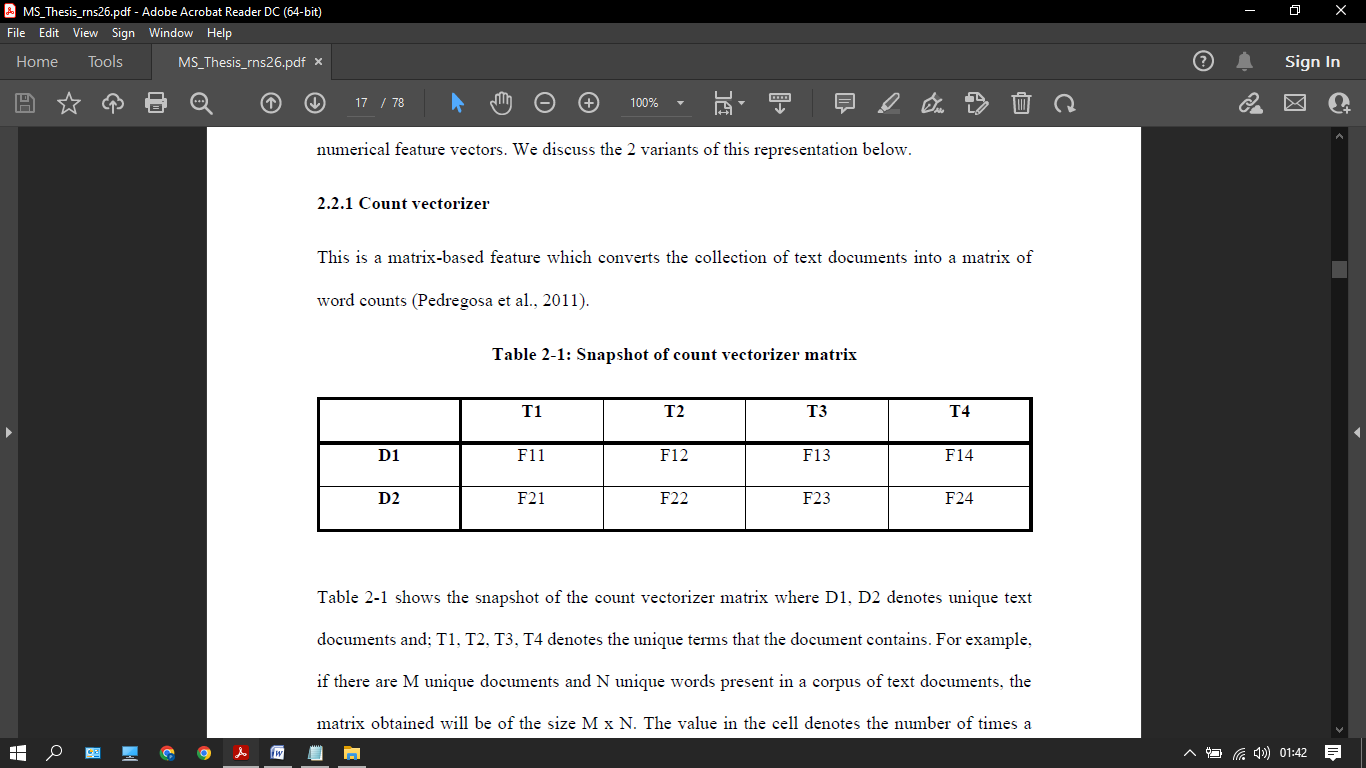
This program that converts a set of textual data to word count matrix.

Figure 1 : Count Vectorizer Matrix

### TF-IDF VECTORIZER

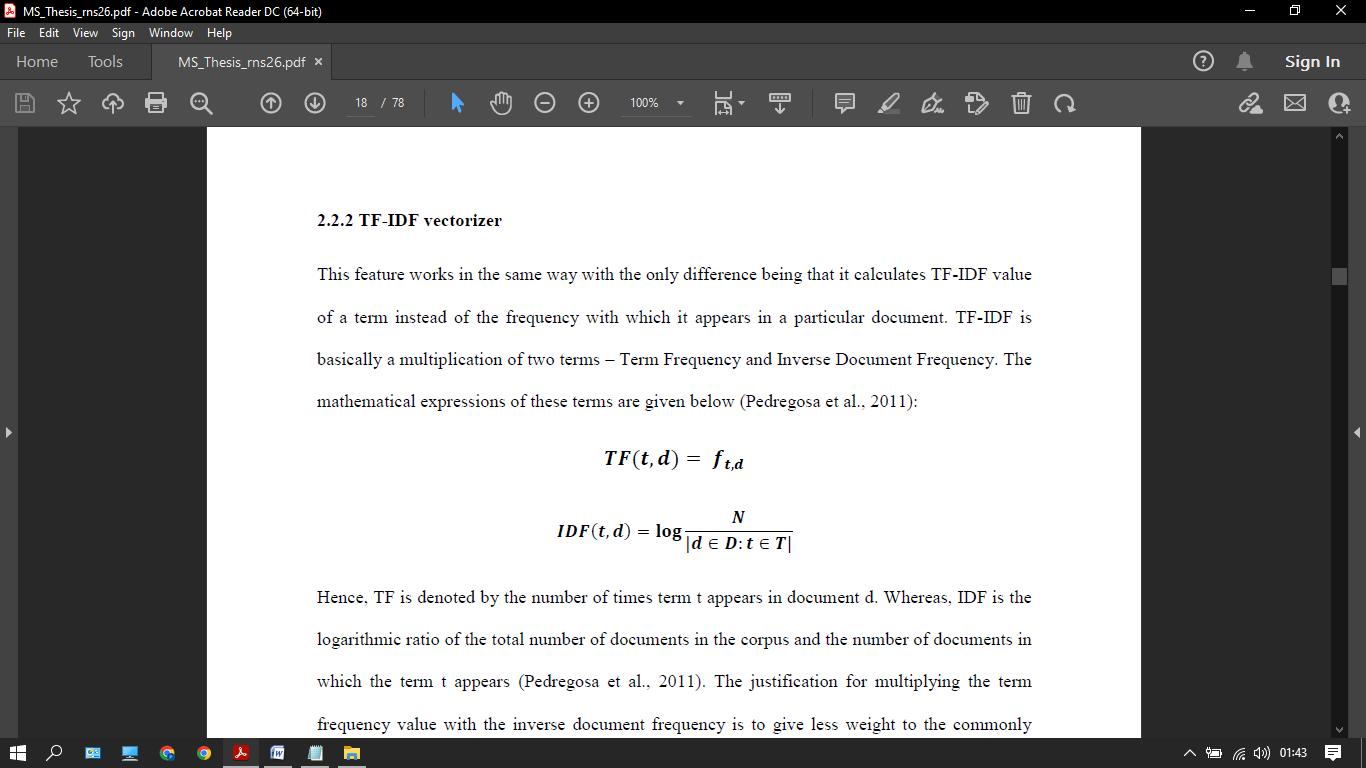


Figure 2: TFIDF Formula

As a result, the number of times t appears in a document d is defined as TF. The IDF, on the other hand, is a logarithmic measures of the overall all docs in a chorus and containing the word t. The term frequency is repeated for a reason. Frequency of the opposite document to provide low priority towards words that appear usually not displays more information.

## 

## SUMMARY FEATURES

The material provided in the review summaries was also assessed in this study, in addition to the review text.

The structural and semantic elements of the summary text are examined. However, because the summary text has less substance than the review text, grammatical characteristics are not taken into account. The text's polarity and subjectivity are both semantic properties.

## ALGORITHMS

### MULTINOMIAL NAÏVE BAYES

This model is a probabilistic method to classify that is based on the Bayes theorem and the naive assumption of feature independence.

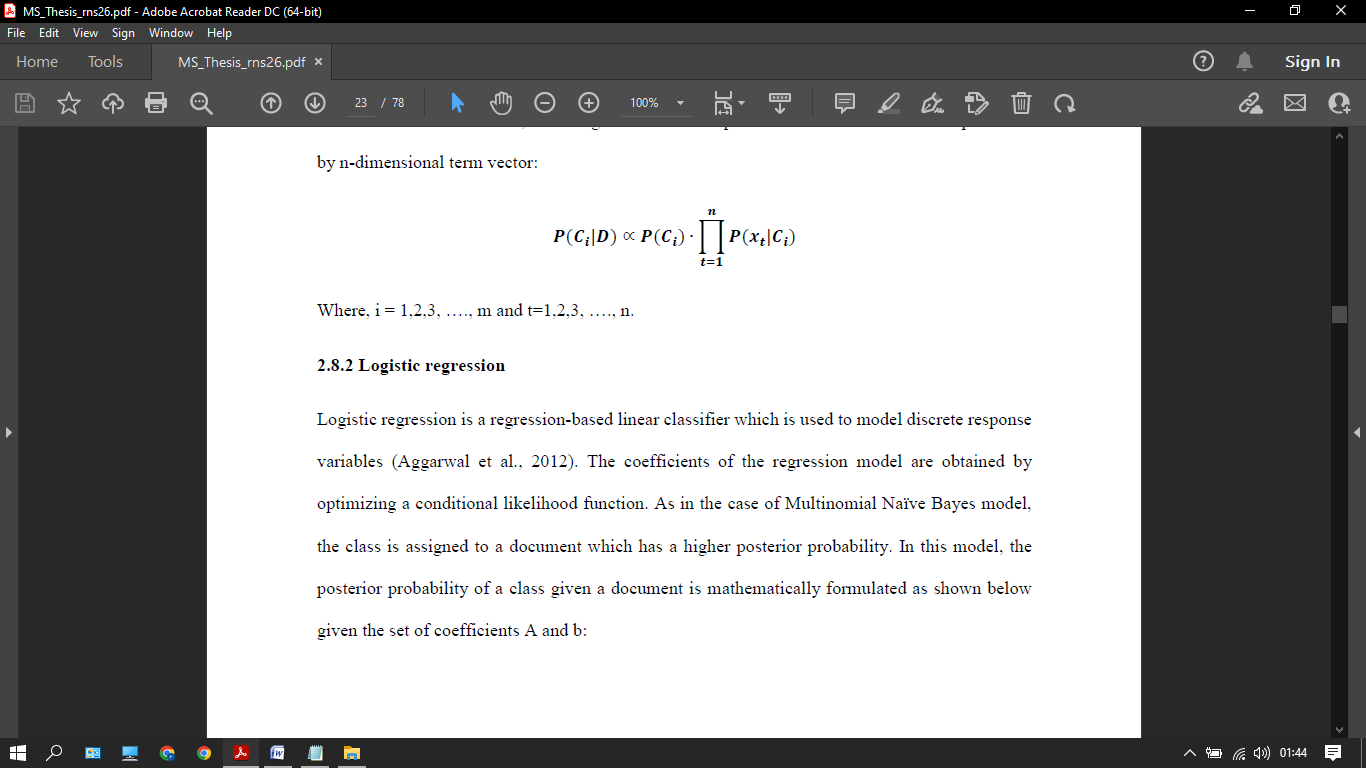


Figure 3: Multinomial Naive Bayes

### DATASET DESCRIPTION

The dataset is available on the Kaggle website. The collection is made up of Amazon.com reviews of exquisite cuisine. It has a total of 500k reviews on 70k products.

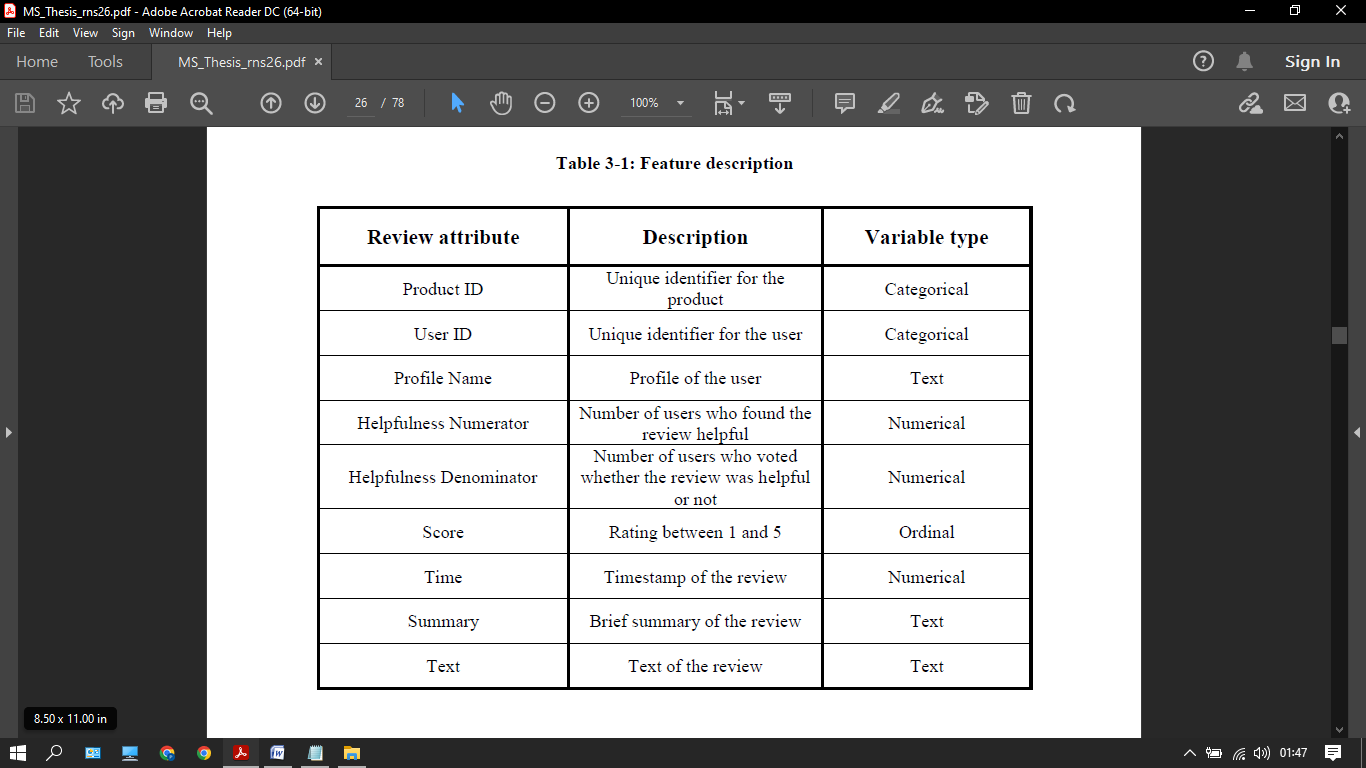


Figure 4: Dataset Description

## 

## METHODOLOGY

This section discusses in detail the process of building an automated text classification system for predicting the helpfulness of online fine food reviews posted on Amazon.com. It encapsulates the following steps:

1. Performing arts searching information analysis for generating the binary response variable
2. Explaining numerous text preprocessing steps that area unit wont to take away creaking terms
3. Explaining the model coaching procedure and also the approaches used for it

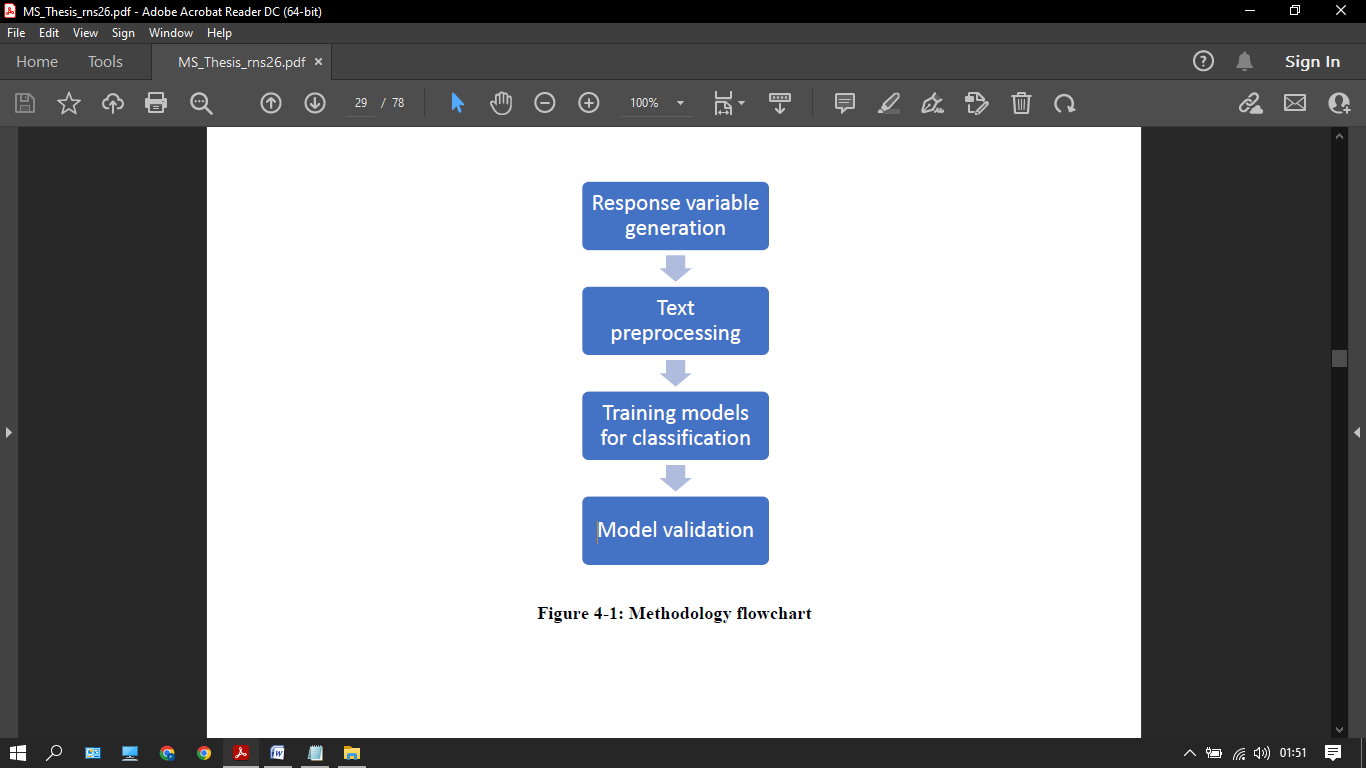


Figure 5: Methodology Flowchart

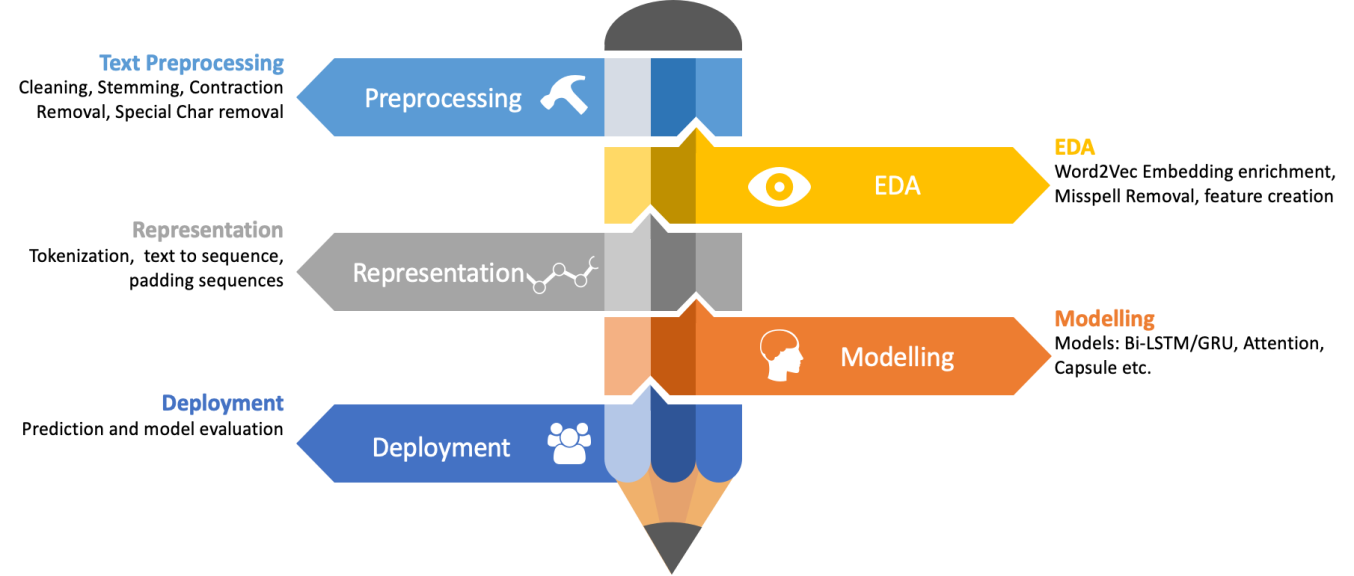


Figure 6: Method

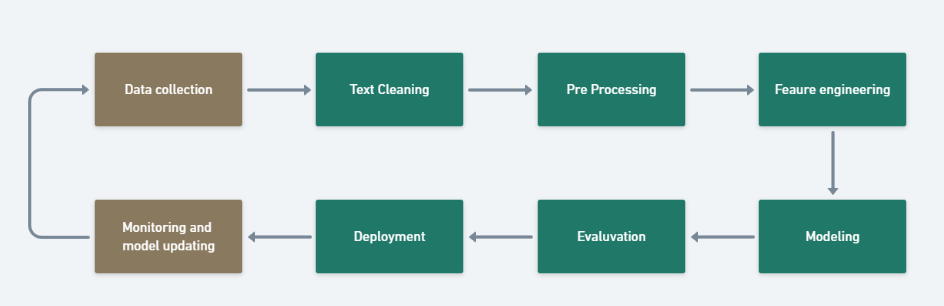


Figure 7: Pipeline

## EXPLORATORY DATA ANALYSIS

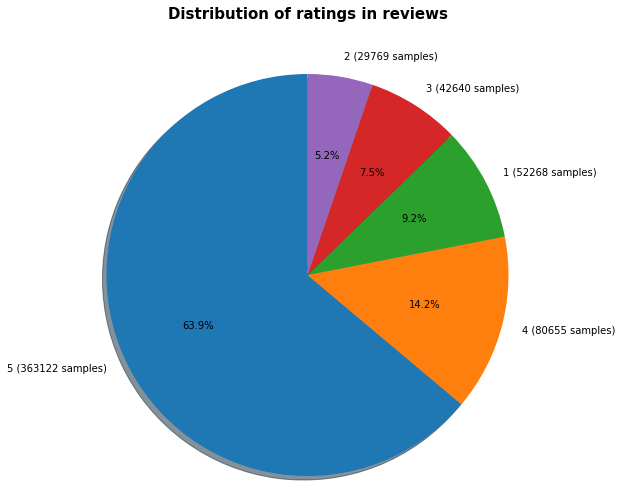


Figure 8: Distribution

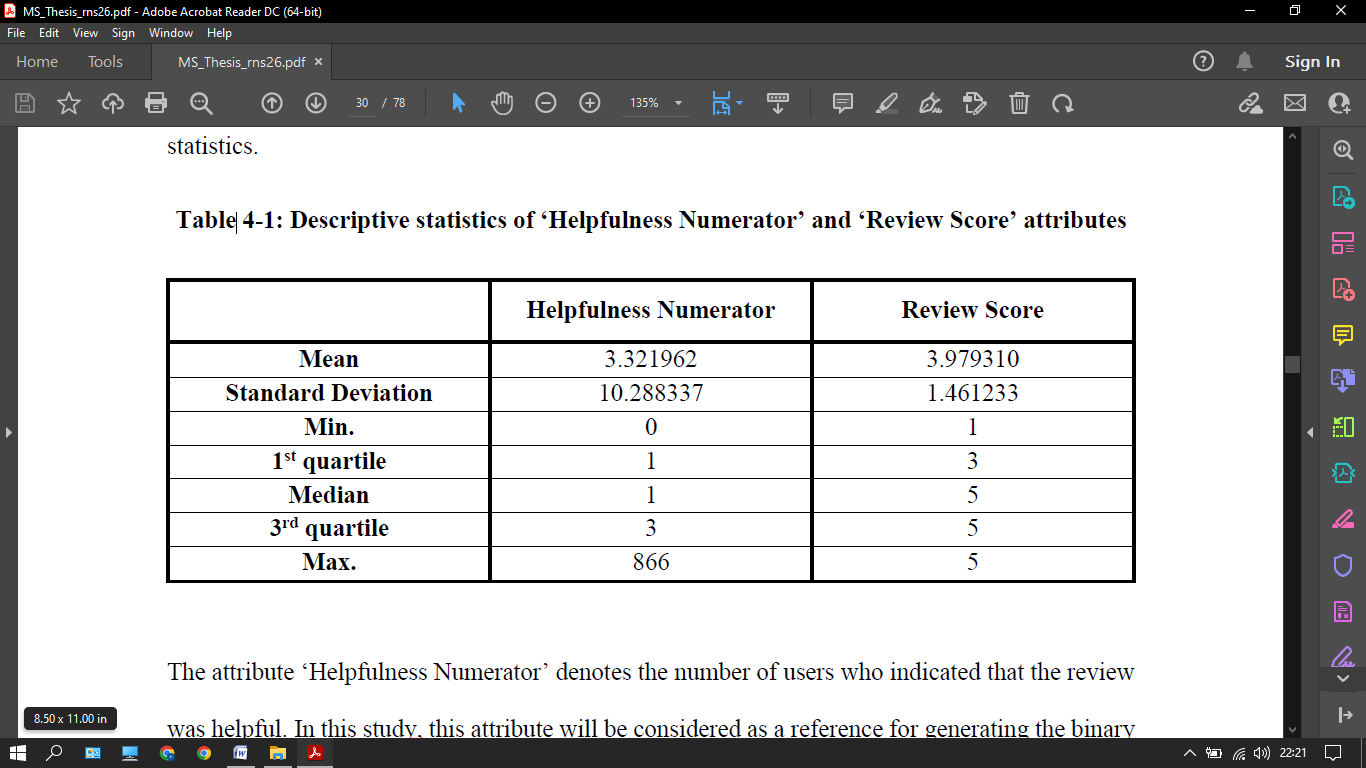


Figure 9: Helpfulness Distribution

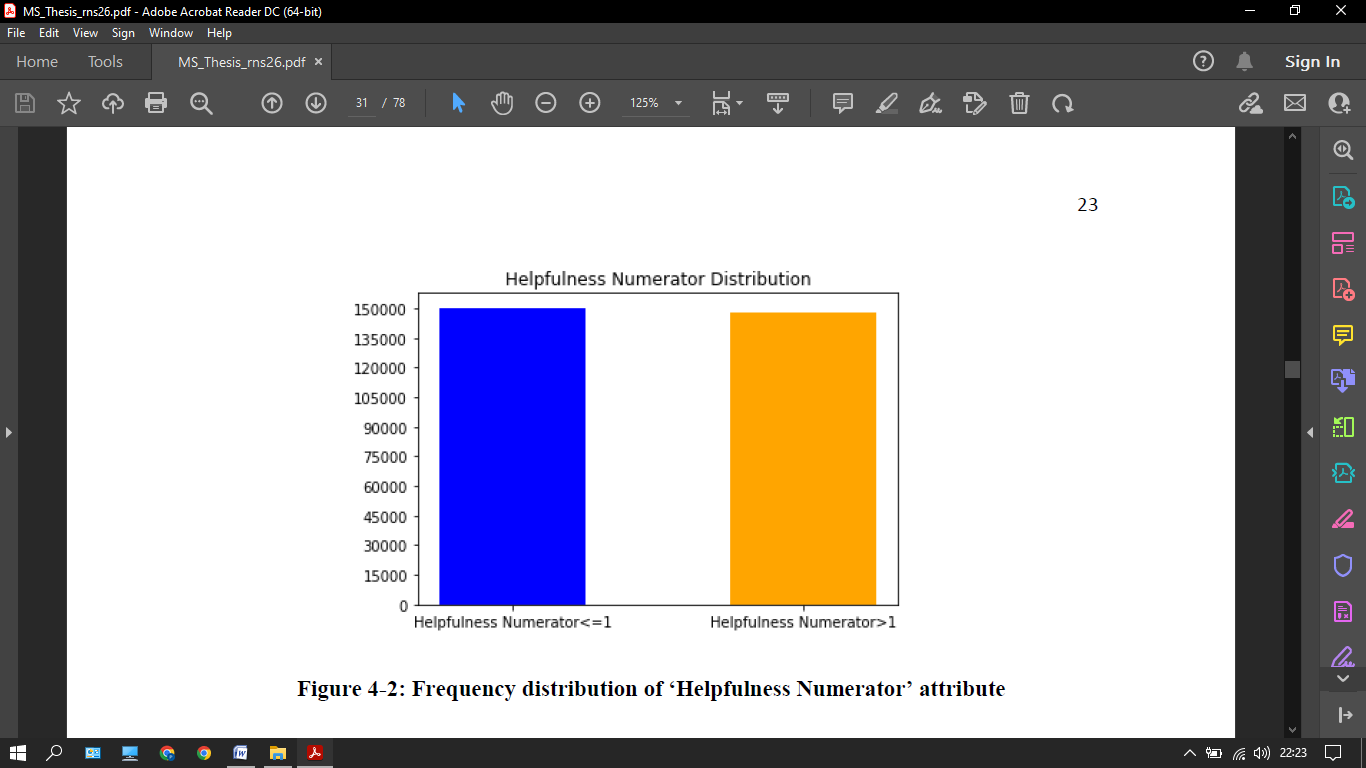


Figure 10: Frequency

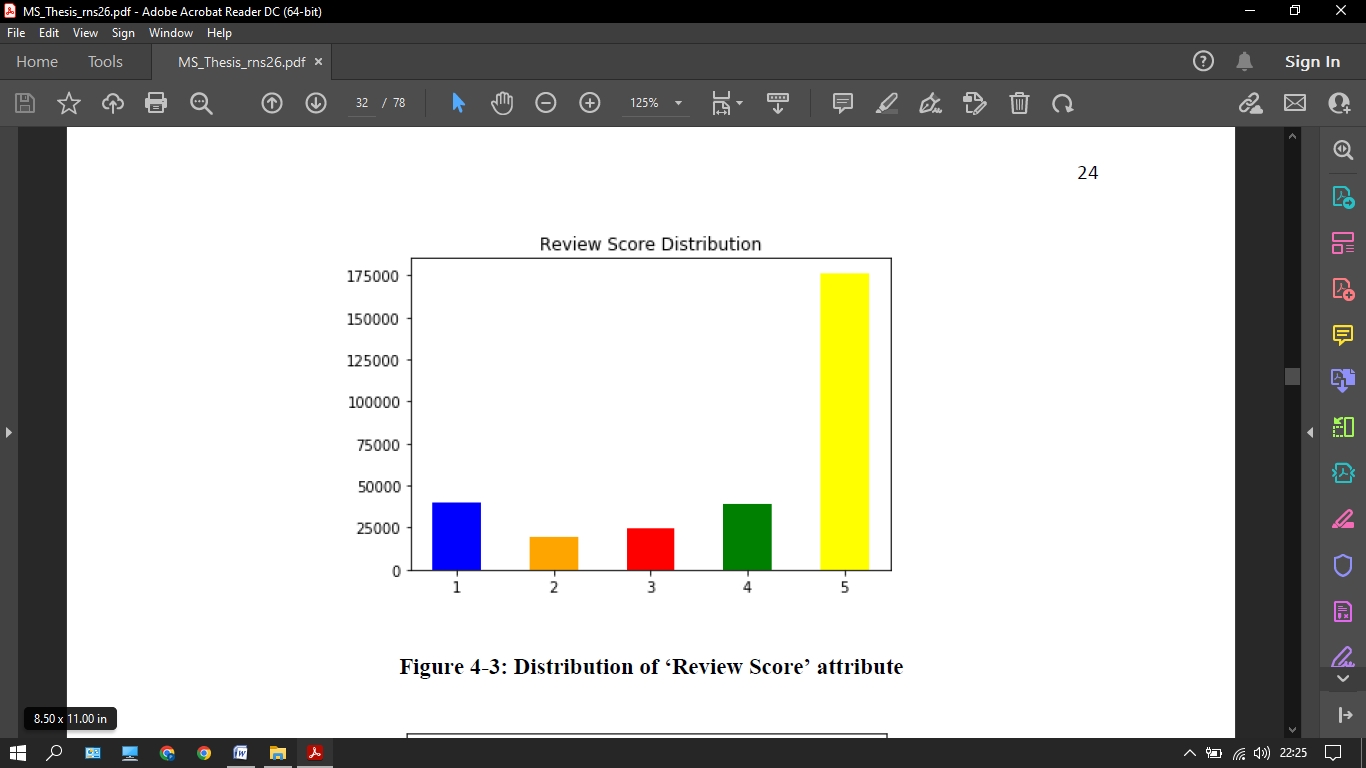


Figure 11: Review Score Distributions

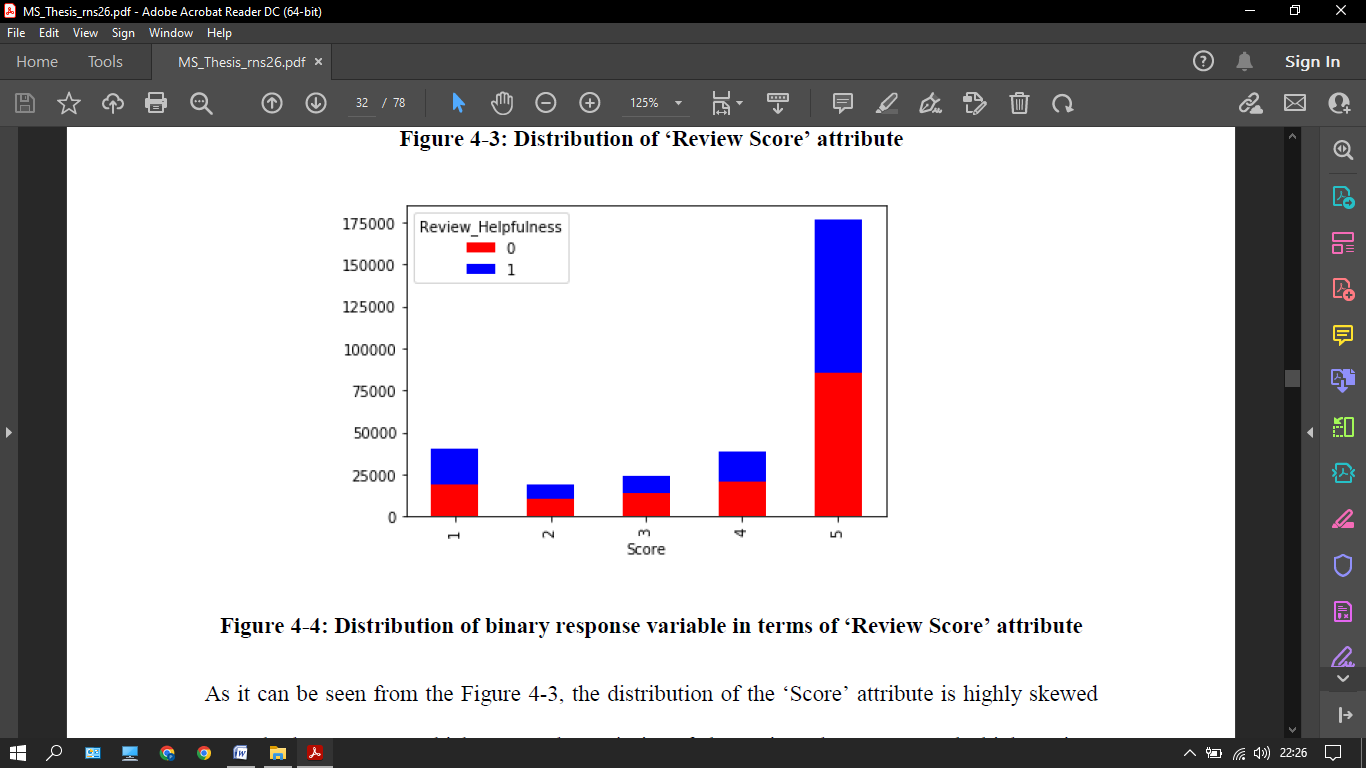
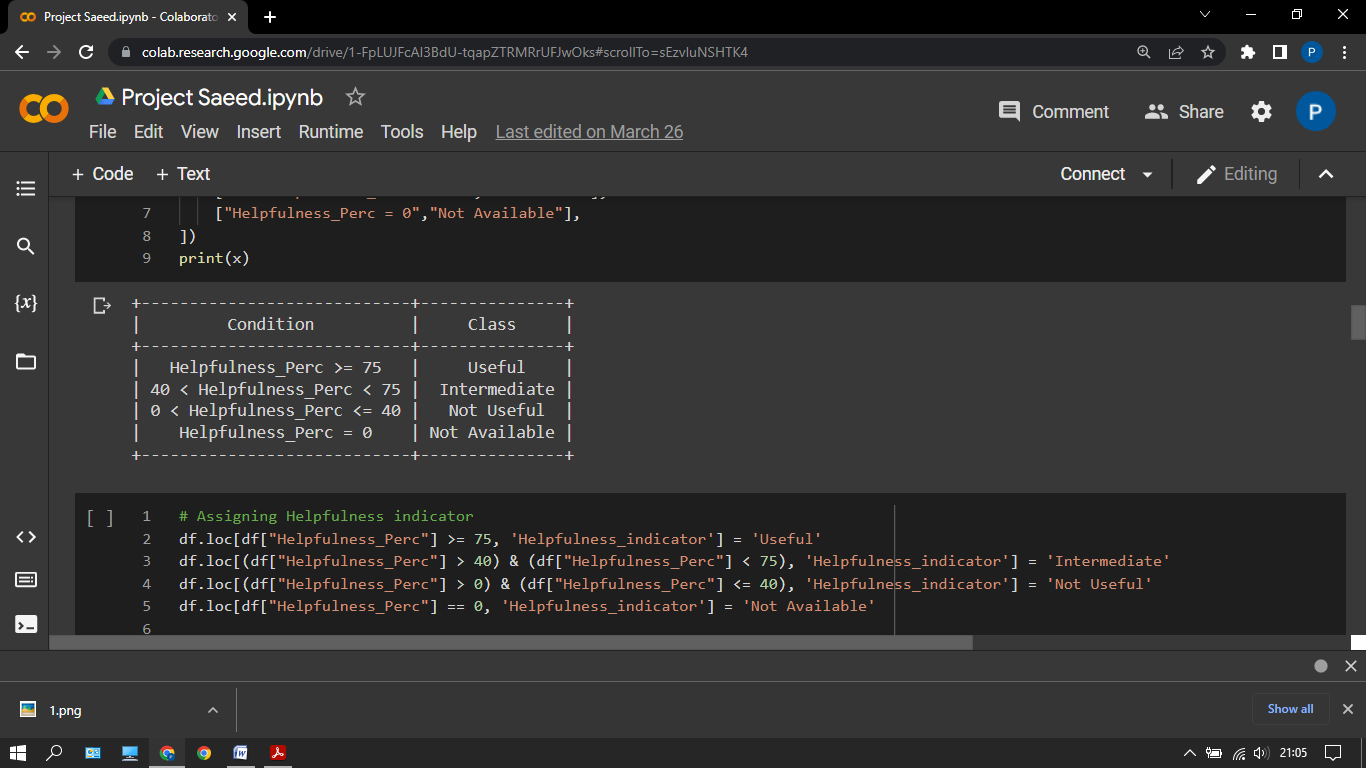


Figure 12: Review distribution



##### Figure 13: Helpfulness Distribution

## 

## TEXT PREPROCESSING

When working with text categorization difficulties, text preparation is critical. By removing the noisy features, it aids in boosting computing performance and avoiding the over-fitting problem.

### 

### TOKENIZATION

This is the 1st move form where we start. It is the process of breaking down phrases into units.

### CONVERTING UPPERCASE TOKENS INTO LOWERCASE

Because the meaning of a word or phrase are unaffected by the case in which it is expressed, uppercase words are changed to lowercase to minimize potential word repetition.

### 

### REMOVAL OF PUNCTUATION MARKS

The punctuation marks are deleted from the text in this stage since they contribute no additional information.

### 

### STOP WORDS

The words like a, an, the, is, are don’t put a major impact on the sentences that is why we removed them.

### LEMMATIZATION

It converts the word into base form of the word from which the future was generated. Like came it changes to come.

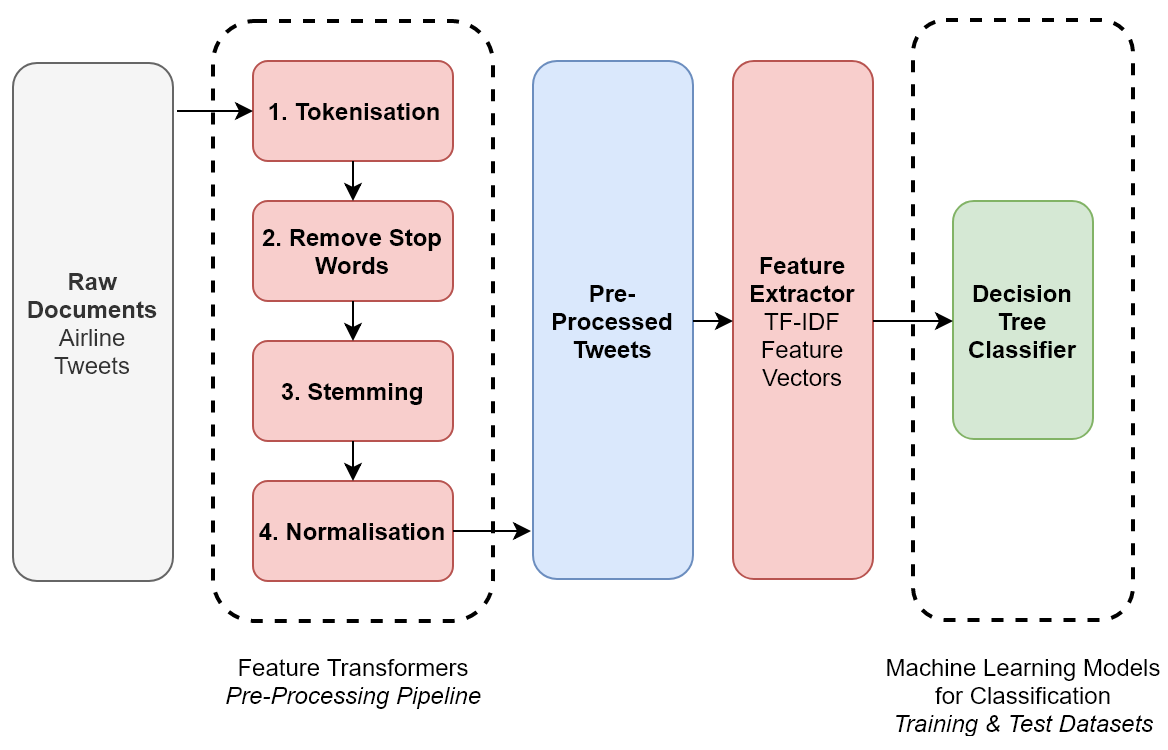


Figure 13: NLP Steps

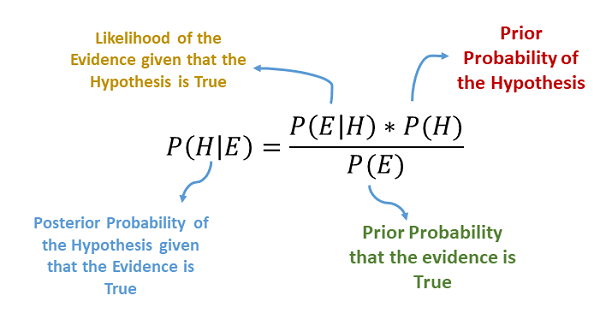


Figure 15: Bayes Rule

## 

## MODEL TRAINING

The review classification job is performed in this work using a variety of methodologies, each of which is paired with a distinct classification algorithm.

This will offer us with a variety of outcome combinations. In addition, each strategy incorporates numerous sets of characteristics that are compared in terms of accuracy.

The vectorized features will be utilized to do classification in the first method.

The document corpus is represented in a matrix format, with each row signifying a document and each column denoting the corpus's unique terms.



## 

## MODEL VALIDATION

For all approaches employed in the study, the database is pushed and divided into 7: 3 train and test sets.

Due to computer constraints, a simpler approach of holding out is favoured over a more sophisticated authentication method such as k-fold authentication.

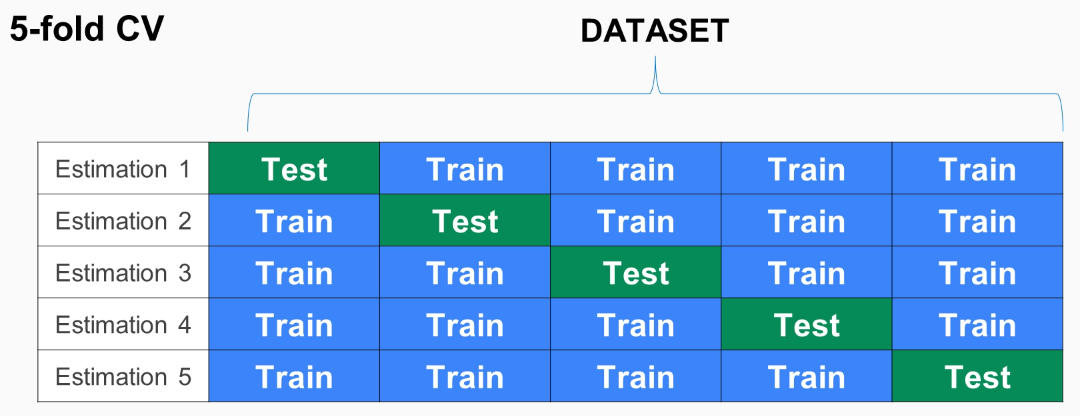


Figure 16: Cross Validation

The catch approach was chosen since the train set's sample size is large enough to prevent the issue of overheating.

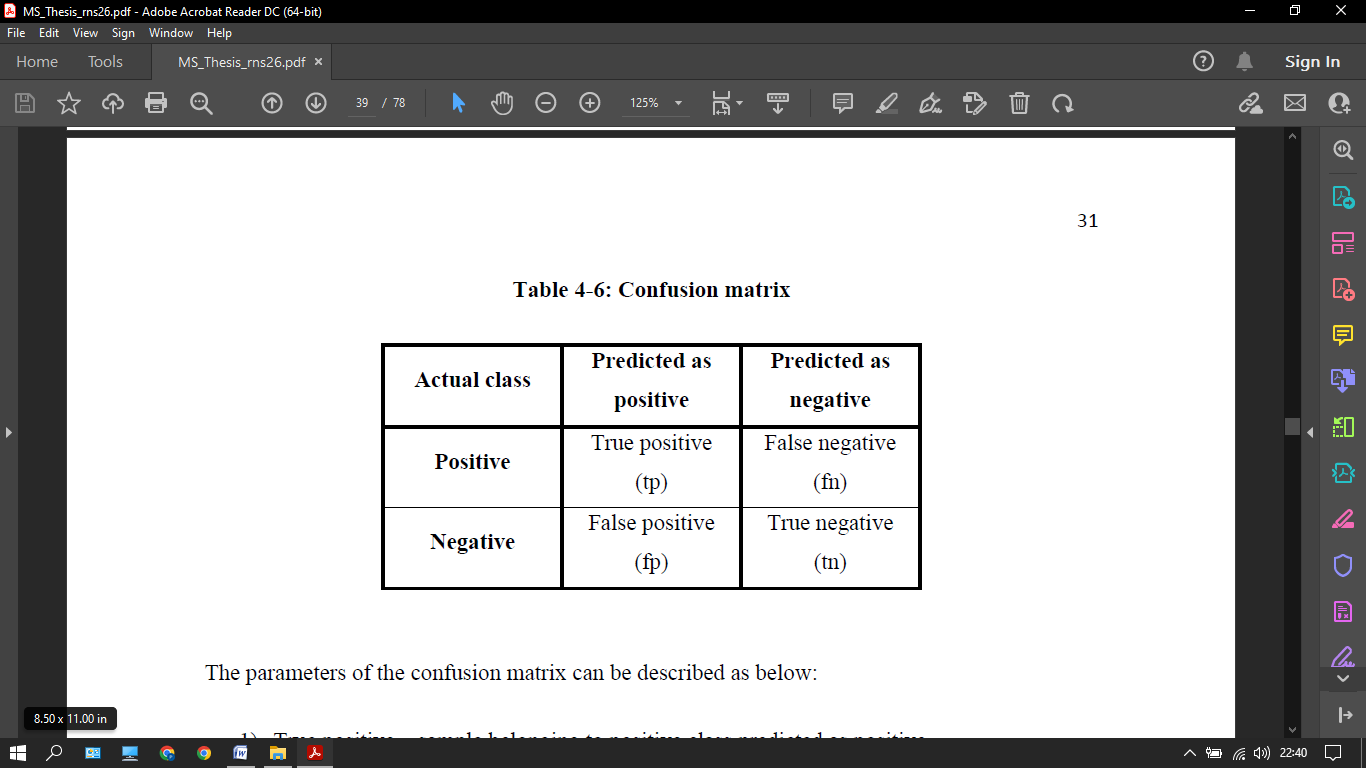


Figure 17: confusion matrix

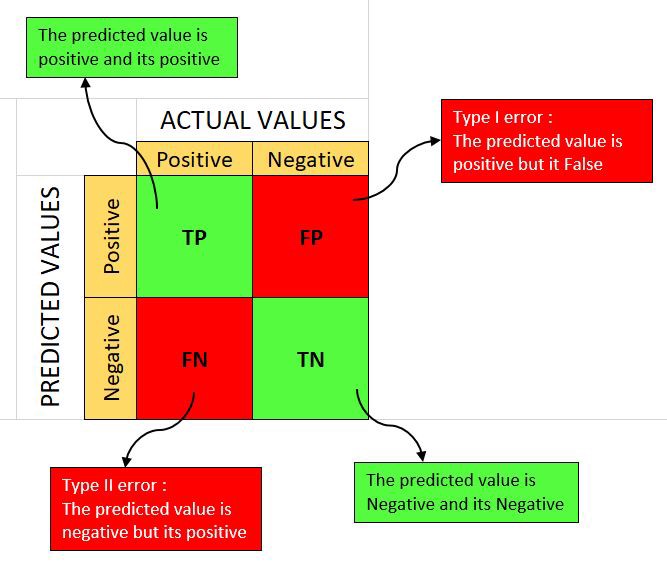


Figure 18: Confusion Matrix

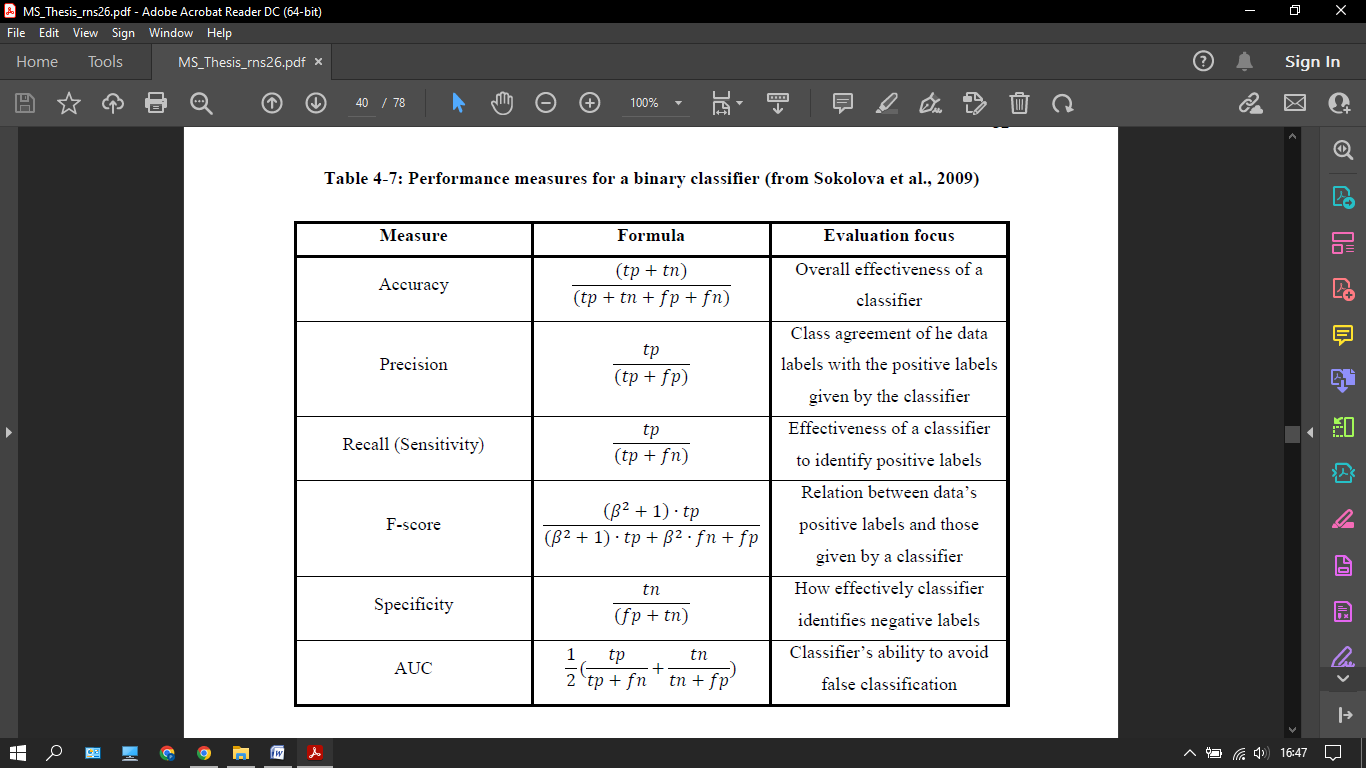


Figure 19: Accuracy All Tables

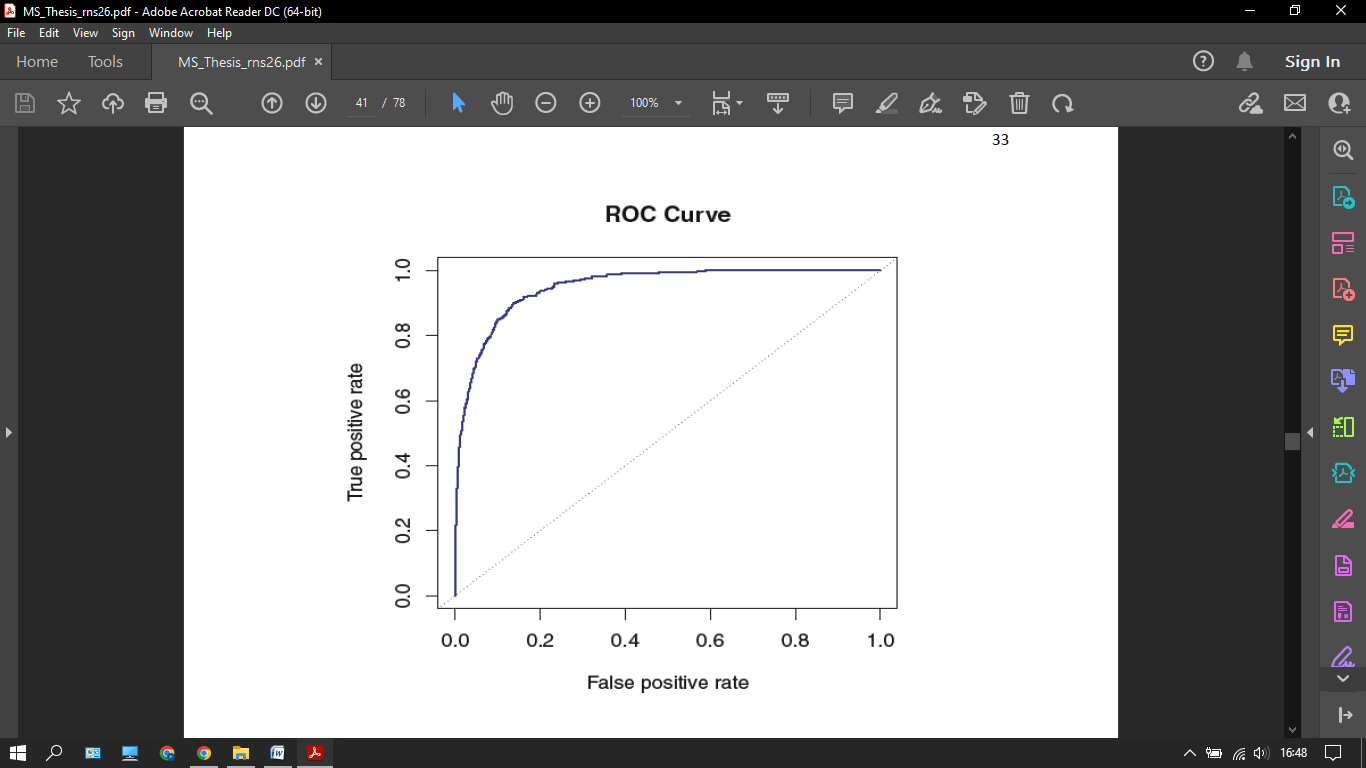


Figure 20: ROC Curve

#### ANALYSIS OF VECTORIZED FEATURES

Here we go through the analysis that was done with the vectorized features. The characteristics are produced in this method by converting the review text into a matrix, with rows denoting unique reviews and columns denoting unique terms in the corpus of review texts. The frequency with which a word appears in a particular review is represented by the element of a count vectorizer matrix, whereas the value obtained by multiplying the term frequency and inverse document frequency values of a word for that particular review is represented by the element of a TFIDF matrix. A single word (unigram) or a pair of adjacent words (bigrams) is also considered as columns of these matrices in this analysis. In this study, both unigrams and a combination of unigrams and bigrams are analyzed and compared in terms of classification performance. The initial phase in the analysis is text preprocessing, which removes noisy words from the matrix and reduces its dimensions. The train and test sets are then generated using random shuffling in a 7:3 ratio. The next step is to create matrices using sci-kit learn to acquire matrix representations of review texts. Naive multinomial by multiplying the posterior probabilities of the words appearing in that class, the Bayes classifier estimates the likelihood of a class given a text

### ANALYSIS OF FEATURES

This section explains how the review and summary centric features, such as structural features, syntactic features, semantic features, and meta-data, were used in the analysis. When it comes to the analysis method, the characteristics for both the review text and the summary text are extracted. For this purpose, observations which do not contain any review summary text are not taken into consideration.

1) Text preprocessing is avoided when extracting structural and syntactic characteristics since it defeats the objective of include them in the research.

2) The objective is to eliminate words that don't help much to understanding the semantics of the review and summary texts, therefore text preparation is done while extracting the semantic characteristics.

**1.)K-NN IN BAG OF WORDS**

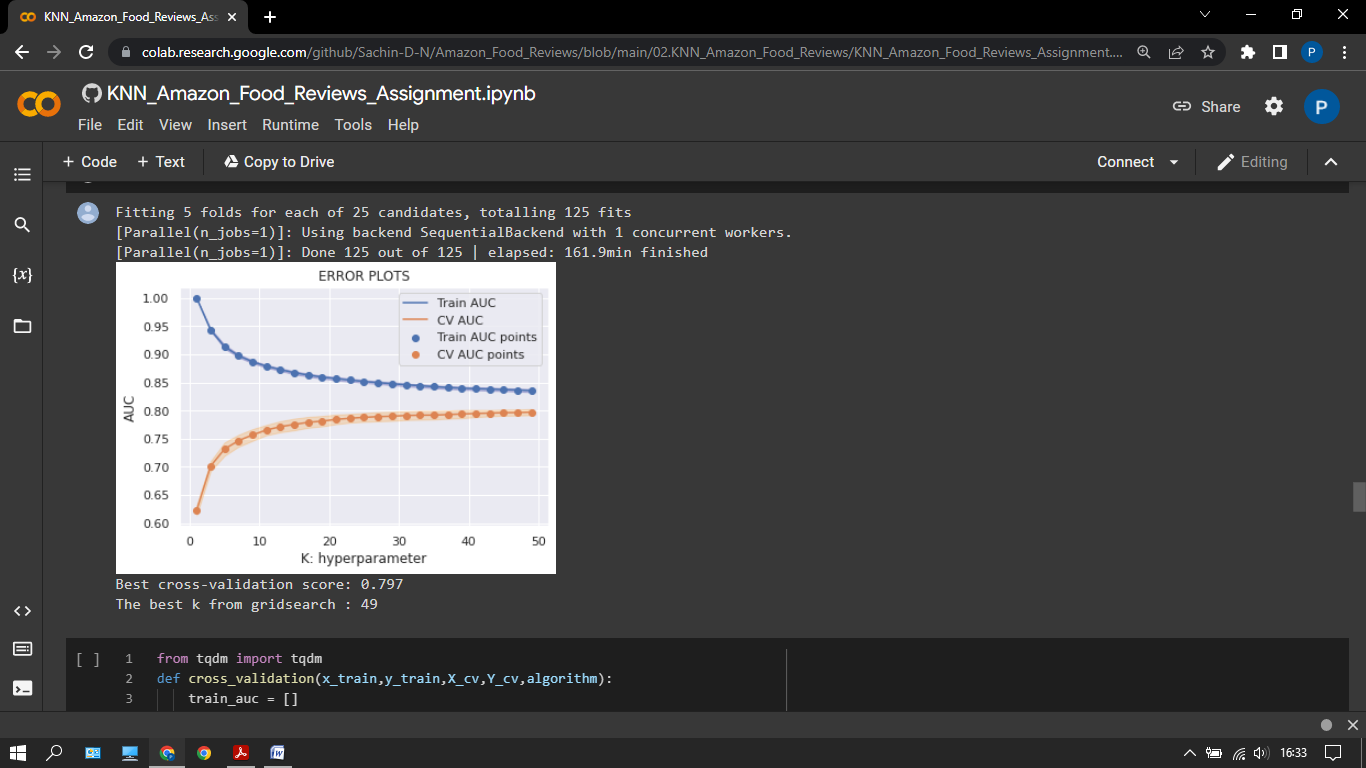


Figure 21: AUC Curve

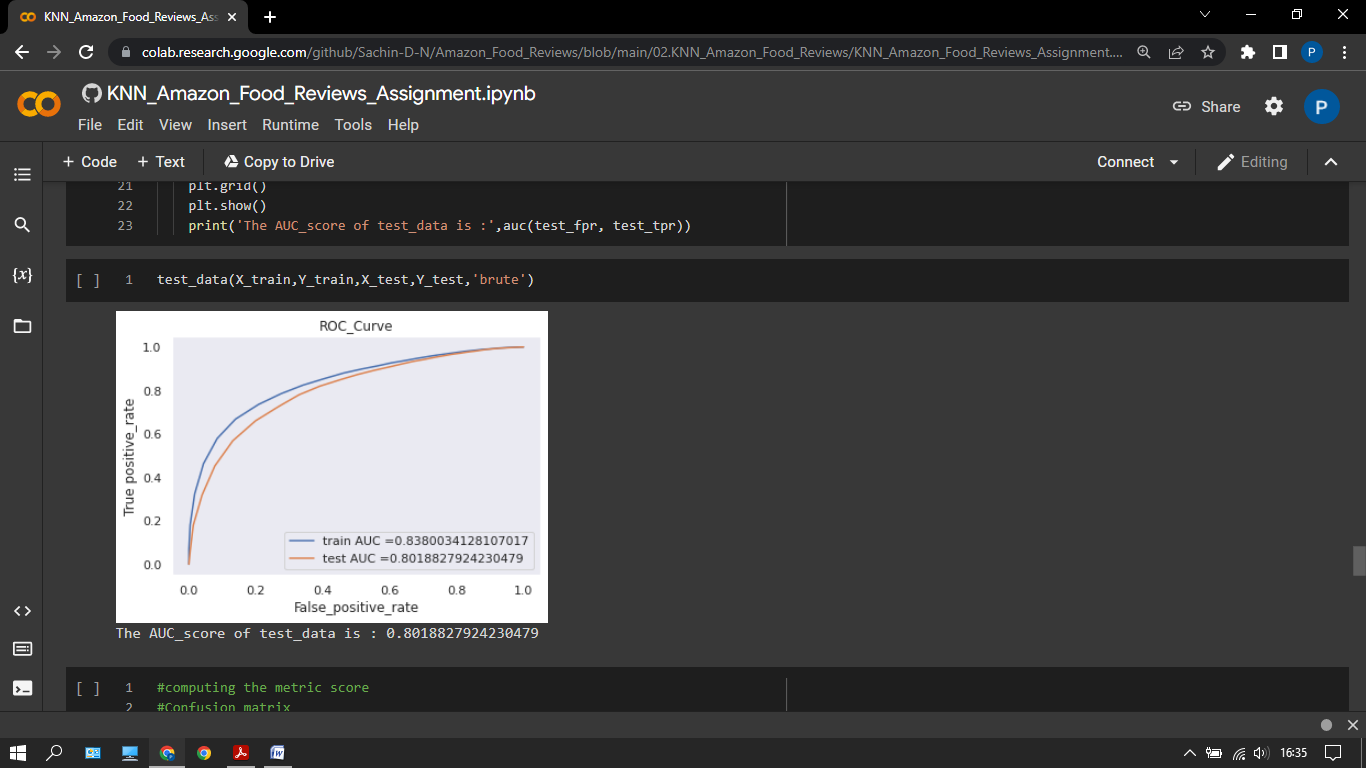


Figure 22: AUC

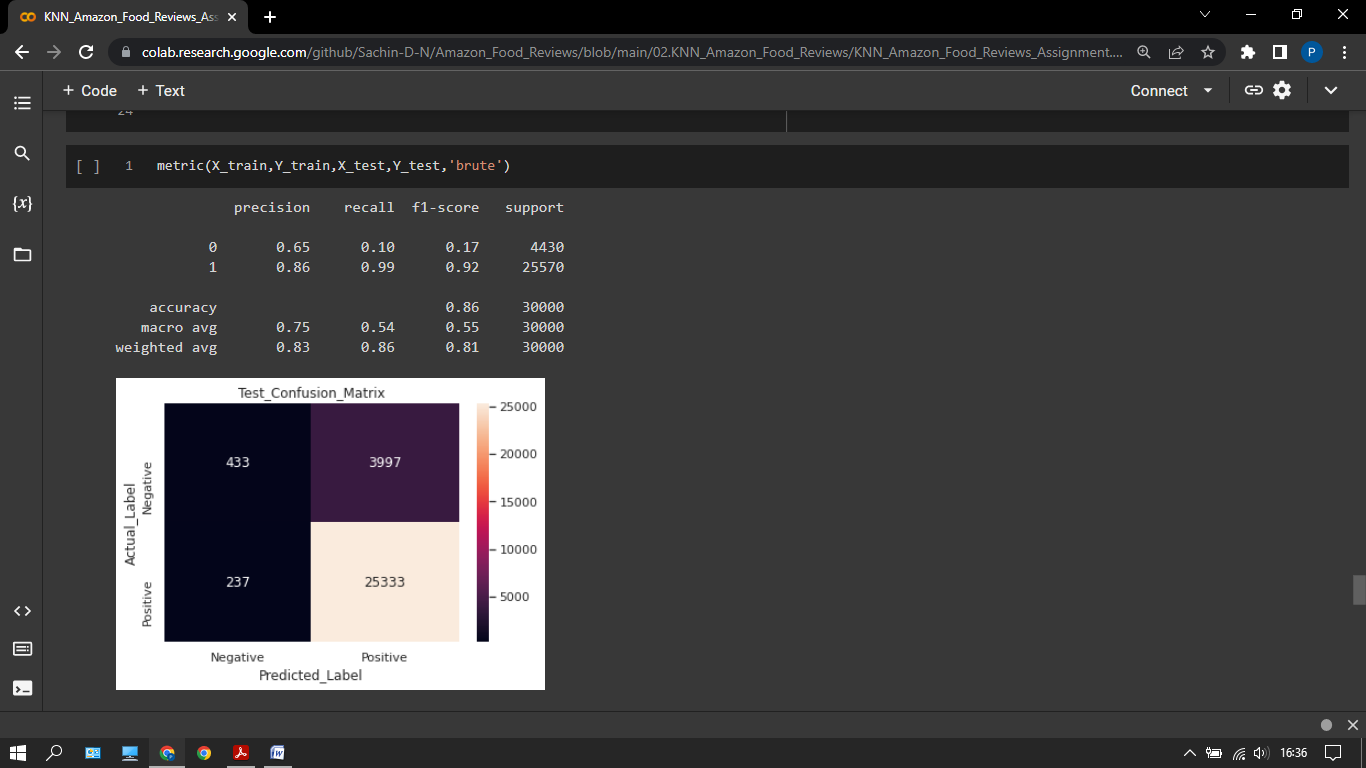


Figure 23: Table

**2.) K-NN IN TFIDF**

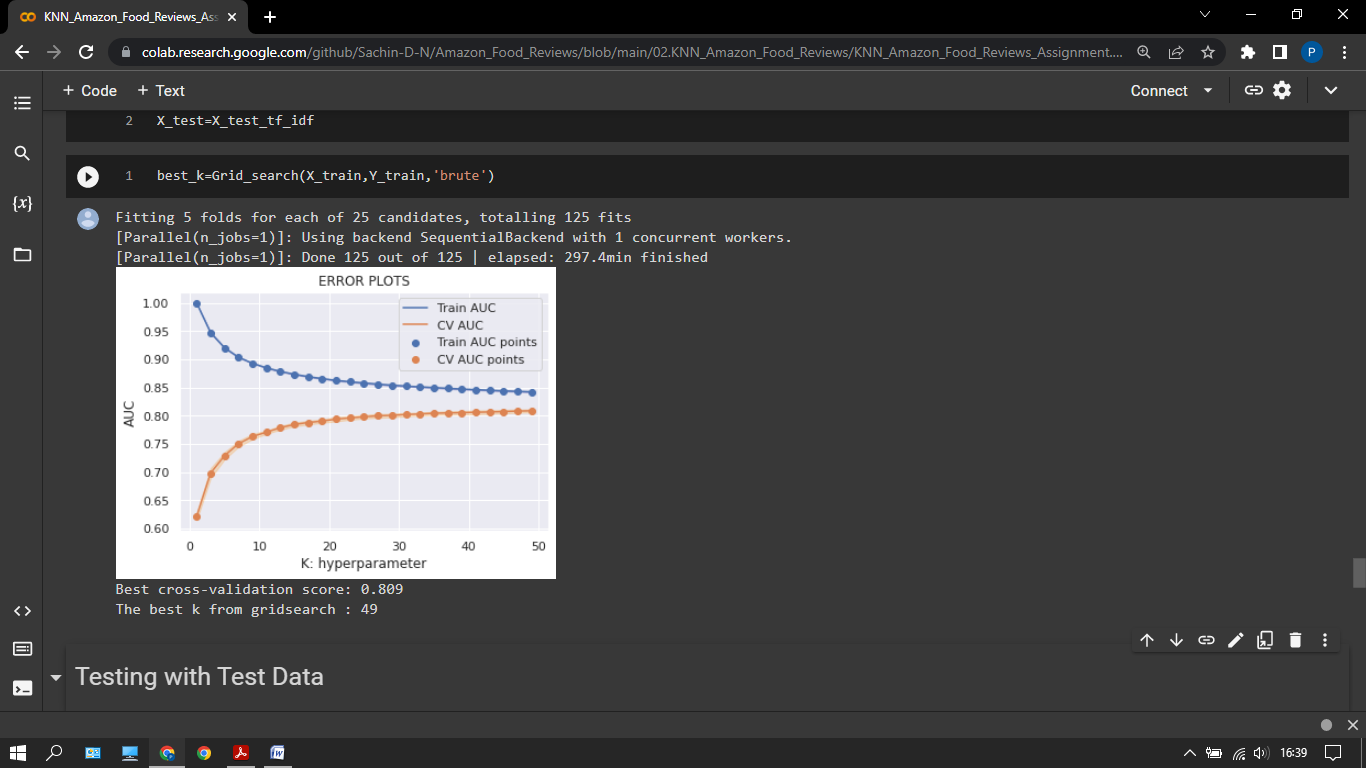


Figure 24: AUC KNN TFIDF

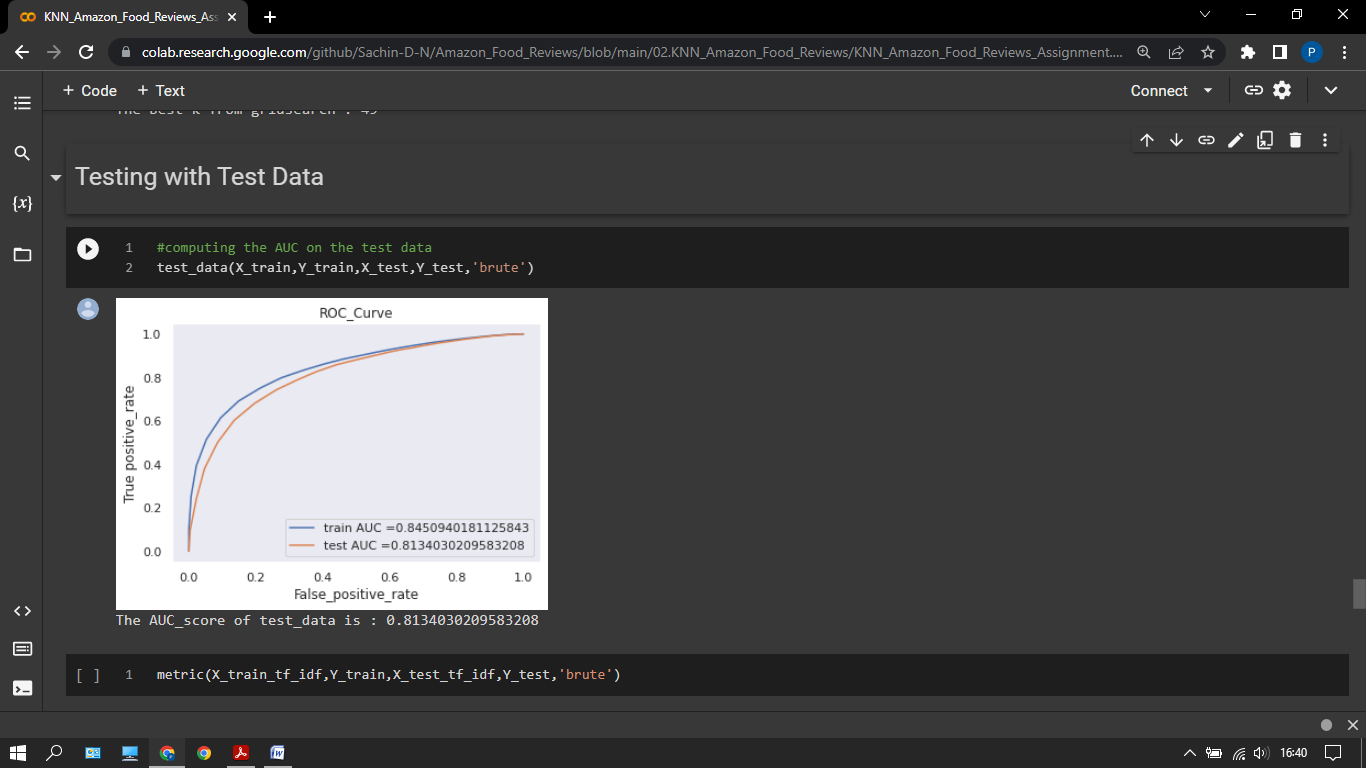


Figure 25: ROC Curve

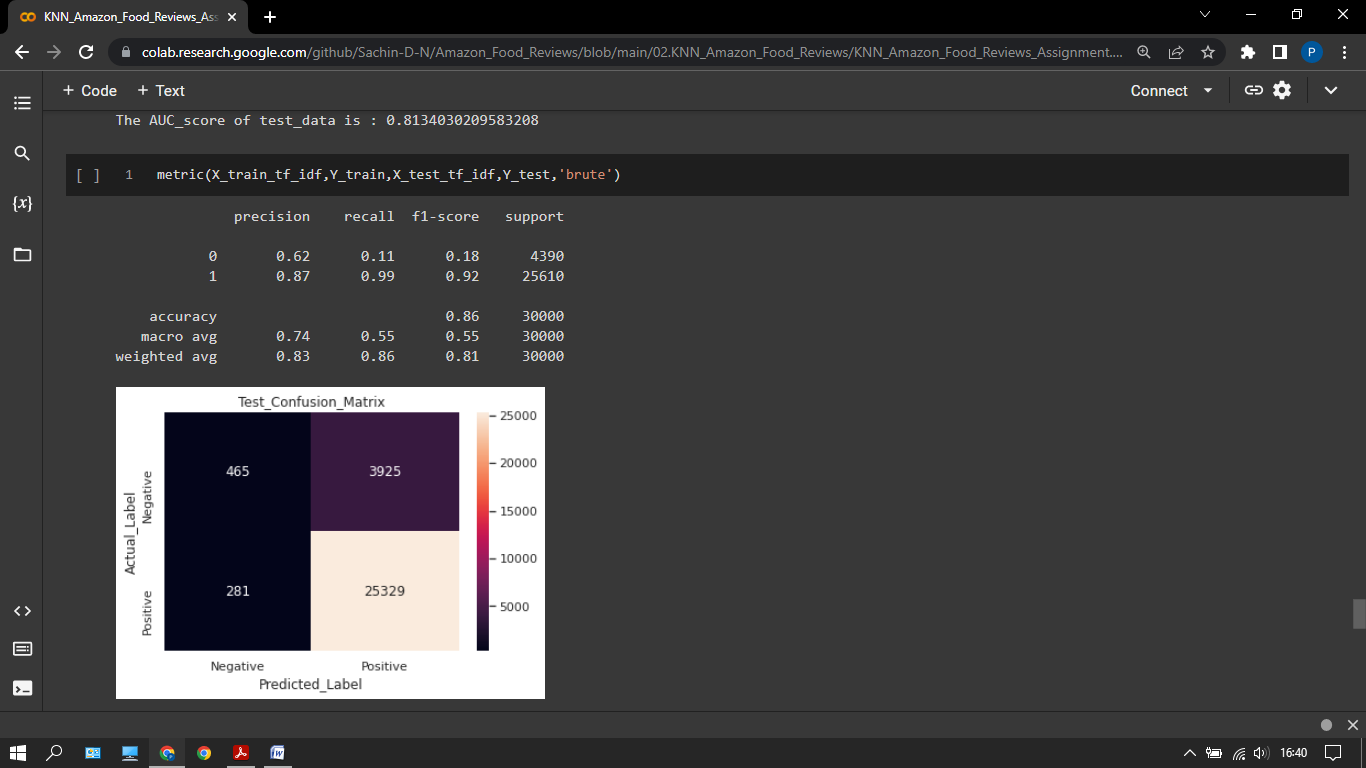


Figure 26: Accuracy

**3.) K-NN IN BAG OF WORDS USING KD-TREE**



Figure 27: AUC KNN

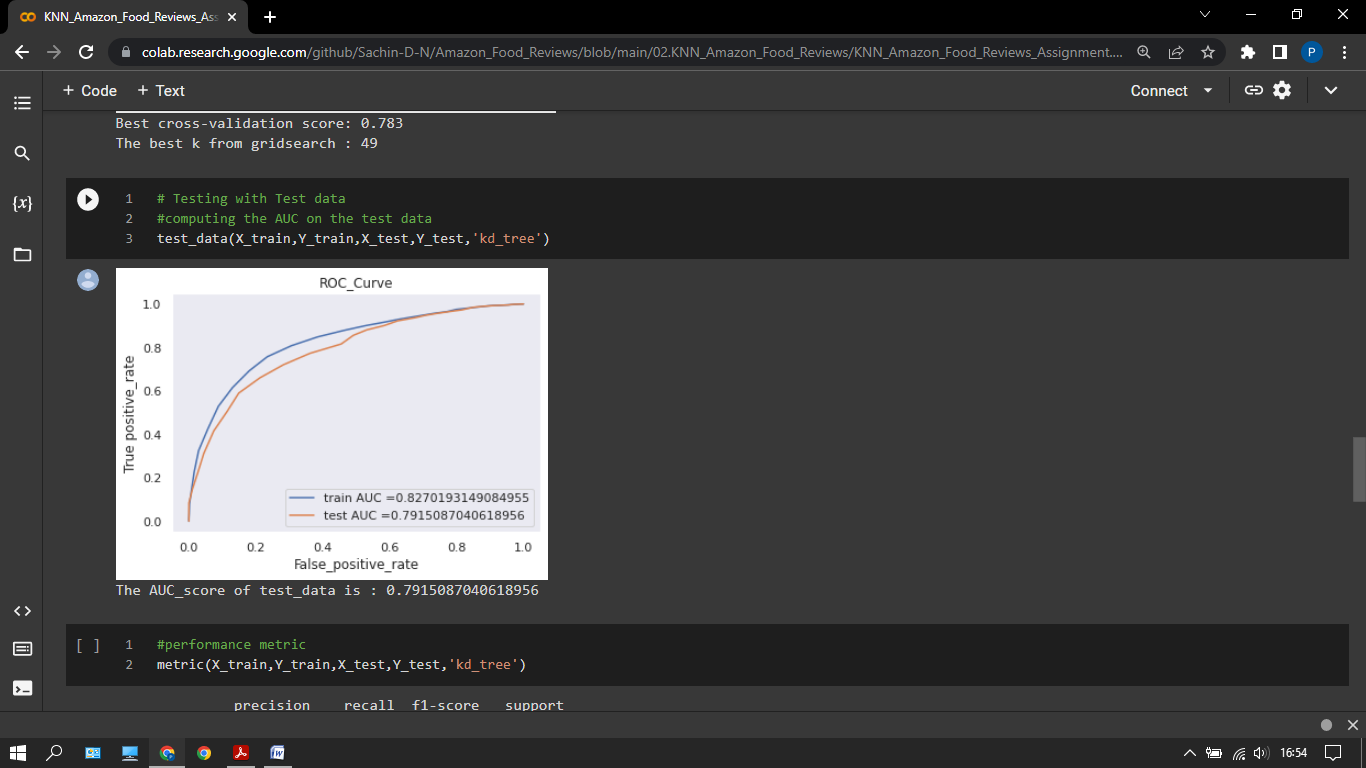


Figure 28: ROC KNN

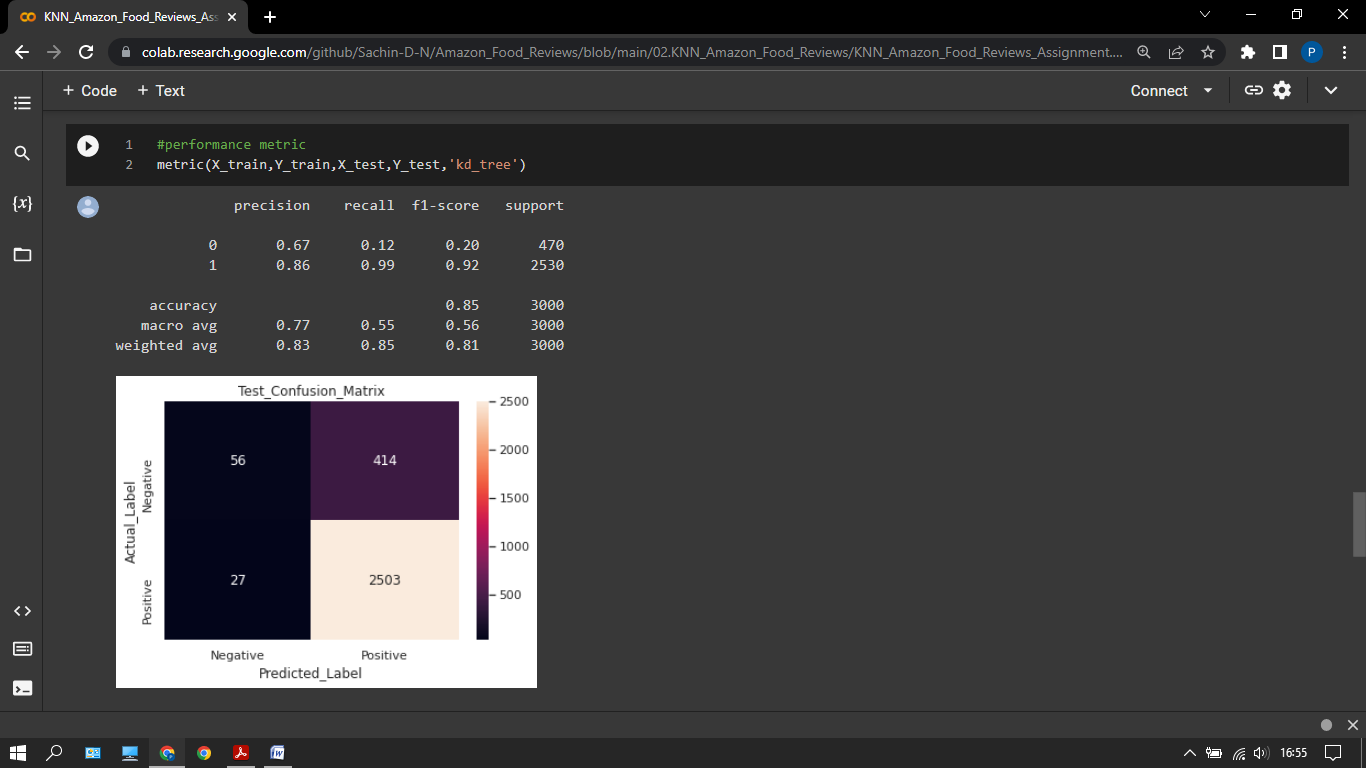


Figure 29: Final KNN

**4.) K-NN WITH KD TREE ON TFIDF**



Figure 30: AUC KNN KD Tree

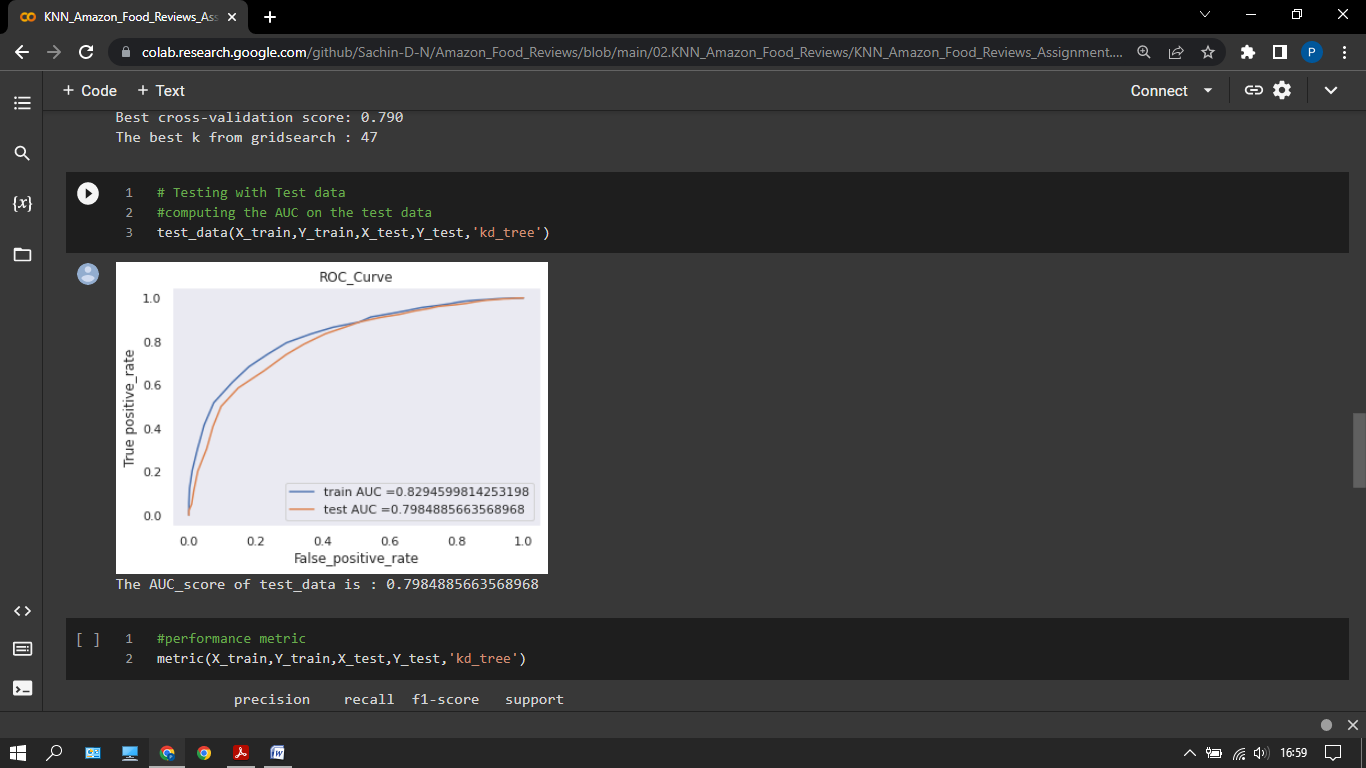


Figure 31: ROC Curve

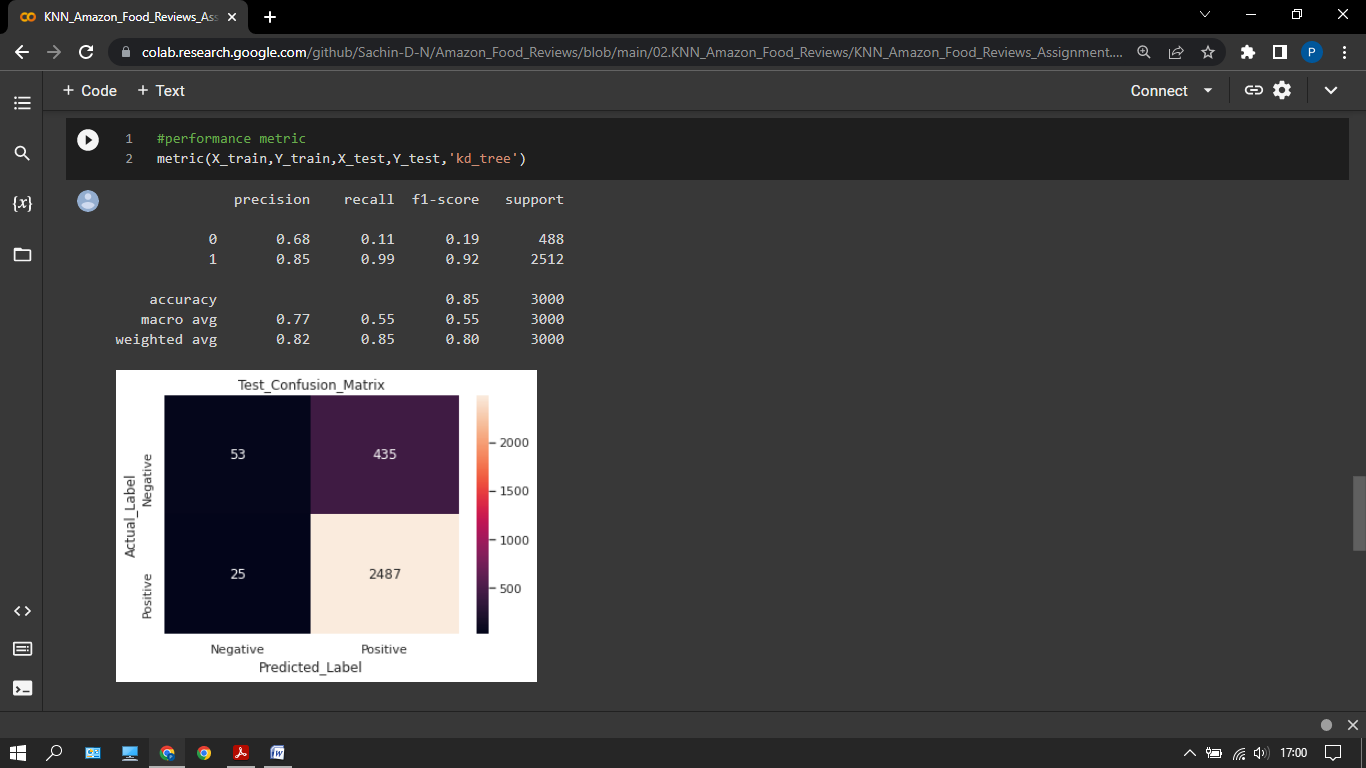


Figure 32: Final knn kdtree

**OBSERVATIONS**

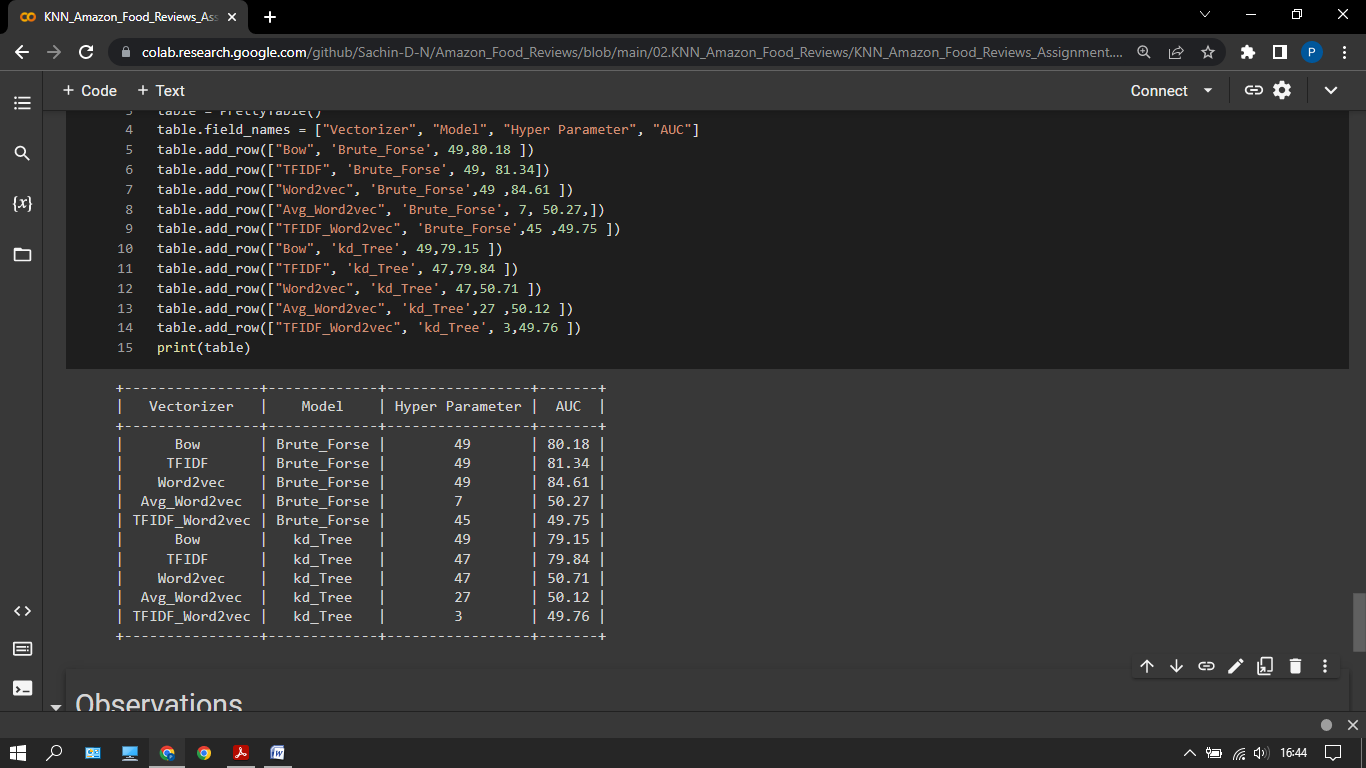


Figure 33: Final Reports

**CONCLUSIONS AND FUTURE WORK**

This research focused on building an automated text classification system that could predict its usefulness regardless of when the online review was published. The goal was to allow both consumers and manufacturers to choose from a wide range of reviews by including the latest unrated reviews in addition to the old, highly rated reviews.

The first step was to conduct a literature review to familiarize you with the work done in the field of text classification, including work related to measuring the usefulness of the review. It turns out that the work on the usefulness of reviews is primarily focused on finding the correlation between the content-oriented characteristics of reviews.

This study performed a comparison between vectorized features, review core features and review summaries, and word embedding-based features. This has never been tried, as far as I know. With the extremely randomized tree, this is a decision tree-based ensemble classifier.

Binary response variables were generated based on the utility votes obtained through the review to perform predictive tasks using the above features and classifiers. Text pre-processing followed, and finally the model was trained and tested using holdout validation along with accuracy and AUC score.

First, you can make the verification procedure more robust. For holdout validation, even if you have enough data to train, you can use the kfold cross-validation to further reduce the chance of over fitting. The vectorized features using Unigram and Bigram achieved the best overall performance. You can also add a trigram to your unigram and bigram combination to see if it improves performance.

Random guessing fared slightly better than features based on word embedding. As a result, a more advanced approach for converting review text into vector form is required.

## 

## REFERENCES

1. Sentiment Analysis: A Definitive Guide , https://monkeylearn.com/
2. Starters Guide to Sentiment Analysis using Natural Language Processing , https://www.analyticsvidhya.com/blog/
3. The basics of NLP and real time sentiment analysis with open source tools , https://towardsdatascience.com/
4. Natural Language Processing and Sentiment Analysis , <https://www.expert.ai/blog/>
5. Natural Language Processing Pipelines , <https://www.kdnuggets.com/2021/03/>
6. NLP Pipeline, <https://towardsdatascience.com/my-first-nlp-pipeline-99d24aafb773>
7. NLP Tutorial , <https://www.javatpoint.com/nlp>
8. Understanding NLP , https://medium.com/analytics-vidhya/understanding-nlp-pipeline-9af8cba78a56
9. Natural Language Toolkit , https://www.nltk.org/
10. Natural Language Processing With Python's NLTK Package ,https://realpython.com/nltk-nlp-python/

## PLAGAIRISM REPORT

