Contents

Chapter 1

Motivation and Literature Review

1.1 Why this Project ?

In the current scenario, technology(social media, internet, smart devices, self-driving cars) is constantly generating loads of textual data every second. As this textual data is cast and unstructured, one can no longer use the common traditional algorithmic approach to understand and derive patterns from this kind of data source; this is where NLP comes into play.

- Natural Language Processing (NLP) is a subfield of artificial intelligence that helps computers understand natural languages. In simple words, NLP is an approach to process, analyze and understand large amounts of text data.

- One can also utilize NLP to gain significant insights into unstructured textual online information using the machines.

- Our goal is to understand various pre-processing and feature extraction techniques for textual data analysis.

1.2 How did you think about this project topic ?

Restaurants play a deciding factor in traveling, and with the increased access to information, new pathways of selecting the best ones emerged.With this dataset, consisting of thousands of reviews crawled from the Tripadvisor website, we can explore “what makes a great restaurant or why people prefer one over another?”

1.3 Three recent related works that mainly involve NLP and feature extraction

- **Twitter sentiment analysis** - This is an automated procedure of identifying and classifying subjective information in text. It can be an opinion/judgment/feeling about a particular topic/discussion, argument or product feature. [1]

- **Hate Speech recognition** - This is an automated procedure of identifying and classifying hate and negative speech.The increasing flow of hate speech on social media and the urgent need for effective counters have drawn significant efforts. Hate speech recognition aims to classify text into non-hate/hate speech. The method may also identify the targeting characteristics like types of hate, such as race, and religion in the text. [2]

- **Pdf to audiobooks** - Using machine learning to transform documents in PDF or image format into audiobooks, using computer vision and text-to-speech.

[1] Related Paper- https://arxiv.org/ftp/arxiv/papers/1601/1601.06971.pdf

[2] Related Paper- https://arxiv.org/pdf/2005.12503v1.pdf

1.4 Literature Review

Sentiment of a text segment can be classified based on users' decision for the particular issue which is obtained from the reviews. The summarization of opinion is the method of finding the aspects which are important for the topic and which are related sentences which are related to reviews due to which the summary can be characterized. There are these modules for the proposed system : summarization, and classification.

- Hotel reviews are retrieved from a website like www.Tripadvisor.in by the technique of web crawling, but we have used a cleaned review dataset from kaggle.

- The reviews are classified into categories by means of machine learning classifiers algorithm .The pre-processed and sentence scores in the text of the classified review are calculated in it.

- Our goal is to understand various pre-processing and feature extraction techniques for textual data analysis.

- The online reviews and evaluations are improved due to this it will be tricky for many customers to differentiate the necessary reviews which can be obtained from the not useful ones. The predictive and descriptive are the two categories that can be separated in the data mining technique. The statistical summarization is nothing but the descriptive mining of the data

- The Sentimental analysis as the technique in it. It will repeatedly collect and extract the sentiments in the sentences. Grammatically the sentences are separated for extracting its support. The presence of the reviews are calculated by the Classification. There are many phases of the reviews such as grammatical errors, spelling etc.

- Classification is a method to classify the data into the assured class based on several similarities or criteria. Classification is additionally called as supervised learning, because of the occurrence square measure given with an illustrious cluster label, a distinction to unsupervised learning during which the label square measure is unidentified.

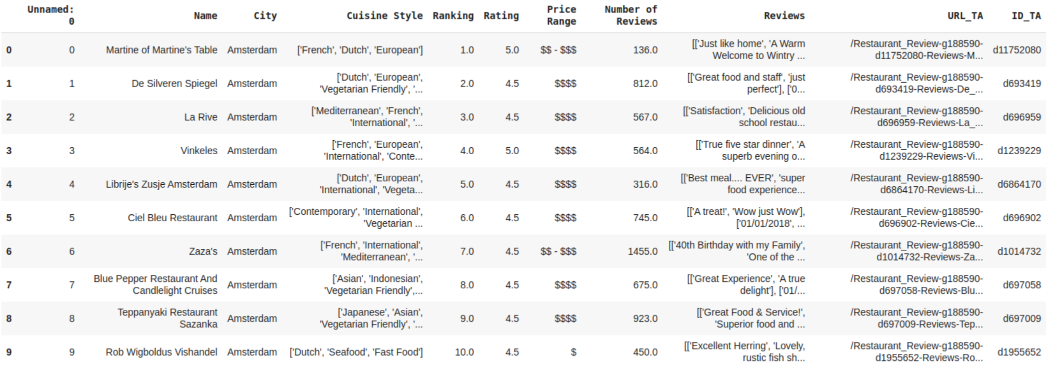
- Classification algorithms are Instance Based for Decision Tree, K-Nearest neighbor, Naive Bayes, Random Forest and Support Vector Machine (SVM) which are compared by means of medium and organization correctness.

Chapter 2

Dataset Description and Preprocessing

2.1 Dataset

Tripadvisor Restaurant review dataset with 11 columns and 125527 ro[1]



	Unnamed: 0	Name	City	Cuisine Style	Ranking	Rating	Price Range	Number of Reviews	Reviews	URL_TA	ID_TA
0	0	Martine of Martine's Table	Amsterdam	['French', 'Dutch', 'European']	1.0	5.0	$$ - $$$	136.0	[['Just like home', 'A Warm Welcome to Wintry ...	/Restaurant_Review-g188590-d11752080-Reviews-M...	d11752080
1	1	De Silveren Spiegel	Amsterdam	['Dutch', 'European', 'Vegetarian Friendly', '...	2.0	4.5	$$$$	812.0	[['Great food and staff', 'just perfect'], ['0...	/Restaurant_Review-g188590-d693419-Reviews-De_...	d693419
2	2	La Rive	Amsterdam	['Mediterranean', 'French', 'International', '...	3.0	4.5	$$$$	567.0	[['Satisfaction', 'Delicious old school restau...	/Restaurant_Review-g188590-d696959-Reviews-La_...	d696959
3	3	Vinkeles	Amsterdam	['French', 'European', 'International', 'Conte...	4.0	5.0	$$$$	564.0	[['True five star dinner', 'A superb evening o...	/Restaurant_Review-g188590-d1239229-Reviews-Vi...	d1239229
4	4	Librije's Zusje Amsterdam	Amsterdam	['Dutch', 'European', 'International', 'Vegeta...	5.0	4.5	$$$$	316.0	[['Best meal.... EVER', 'super food experience...	/Restaurant_Review-g188590-d6864170-Reviews-Li...	d6864170
5	5	Ciel Bleu Restaurant	Amsterdam	['Contemporary', 'International', 'Vegetarian ...	6.0	4.5	$$$$	745.0	[['A treat!', 'Wow just Wow'], ['01/01/2018', ...	/Restaurant_Review-g188590-d696902-Reviews-Cie...	d696902
6	6	Zaza's	Amsterdam	['French', 'International', 'Mediterranean', '...	7.0	4.5	$$ - $$$	1455.0	[['40th Birthday with my Family', 'One of the ...	/Restaurant_Review-g188590-d1014732-Reviews-Za...	d1014732
7	7	Blue Pepper Restaurant And Candlelight Cruises	Amsterdam	['Asian', 'Indonesian', 'Vegetarian Friendly',...	8.0	4.5	$$$$	675.0	[['Great Experience', 'A true delight'], ['01/...	/Restaurant_Review-g188590-d697058-Reviews-Blu...	d697058
8	8	Teppanyaki Restaurant Sazanka	Amsterdam	['Japanese', 'Asian', 'Vegetarian Friendly', '...	9.0	4.5	$$$$	923.0	[['Great Food & Service!', 'Superior food and ...	/Restaurant_Review-g188590-d697009-Reviews-Tep...	d697009
9	9	Rob Wigboldus Vishandel	Amsterdam	['Dutch', 'Seafood', 'Fast Food']	10.0	4.5	$	450.0	[['Excellent Herring', 'Lovely, rustic fish sh...	/Restaurant_Review-g188590-d1955652-Reviews-Ro...	d1955652

Figure 2.1: Dataset

- **Numerical Columns -** Ranking, Rating, Number of Reviews
- **Categorical COlumns -** Name, City, Cuisine Style, Price Range, Reviews, URL-TA, ID-TA

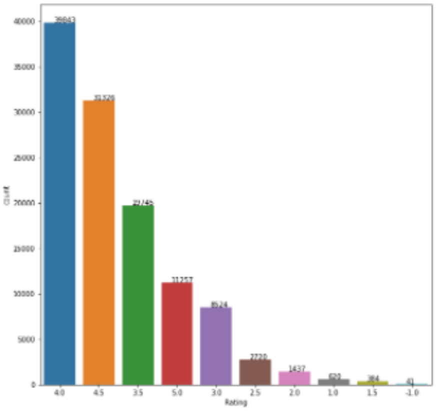
[1] Data Source- https://www.kaggle.com/damienbeneschi/krakow-ta-restaurans-data-raw

2.2 Exploratory Data Analysis

As we are more focused on Reviews and Rating features, the project will mostly contain the steps which are given below

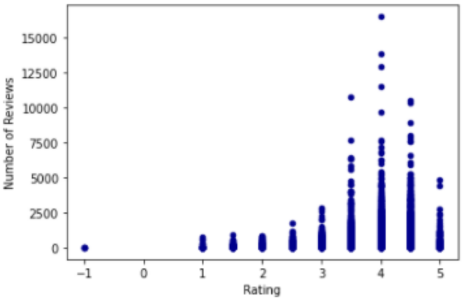
Histogram of Rating feature

The X-axis has Ratings in descending order. Data is an imbalance in regard to the number of restaurants belonging to a particular Rating



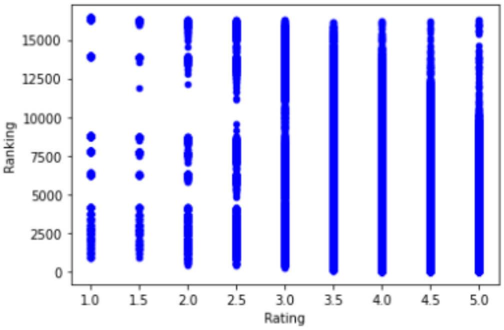
Number of Reviews VS Rating

-1 Rating has no Reviews.



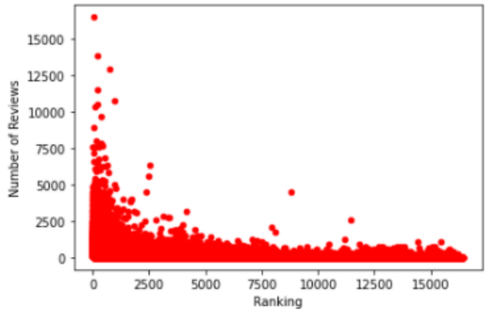
Rating VS Ranking

Almost all restaurants throughout the rank have received all categories of Rating.



Ranking VS Number of Reviews

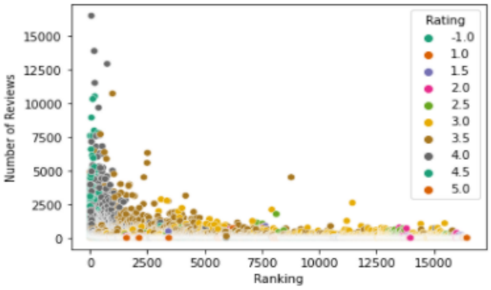
Reviews are written for almost all ranked Restaurants, but lower Ranked restaurants have more Rating, which can also be the reason for human psychology that a restaurant with good rank and more reviews will be good.



Ranking VS Number of Reviews with Rating as categories

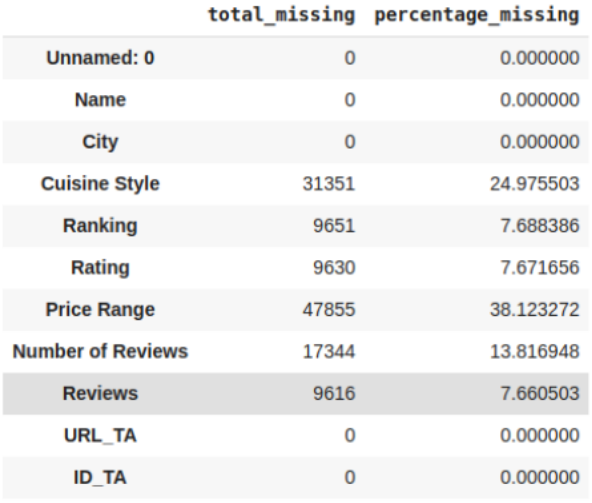
All Rating spread over all Ranked restaurants.

.



Missing Data

Six columns have missing values



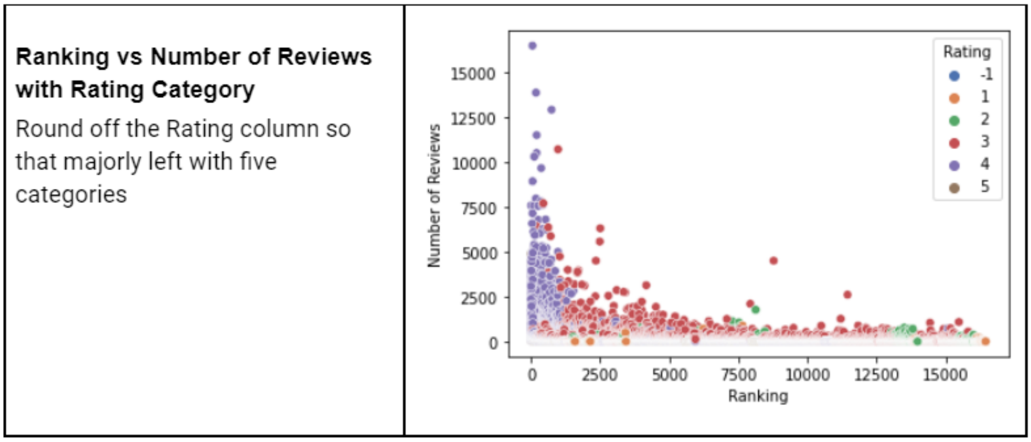
	total_missing	percentage_missing
Unnamed: 0	0	0.000000
Name	0	0.000000
City	0	0.000000
Cuisine Style	31351	24.975503
Ranking	9651	7.688386
Rating	9630	7.671656
Price Range	47855	38.123272
Number of Reviews	17344	13.816948
Reviews	9616	7.660503
URL_TA	0	0.000000
ID_TA	0	0.000000

2.3 Pre-Processing

- **Removing Columns** - 'URL-TA', 'ID-TA' ,'Price Range' all these three features are removed as these all are extra features for us that no usage for our model along with the first 'Unnamed: 0' column that was initially index of the dataset but as pandas by default handle indexing we removed it. New Shape of the data (125527, 7)

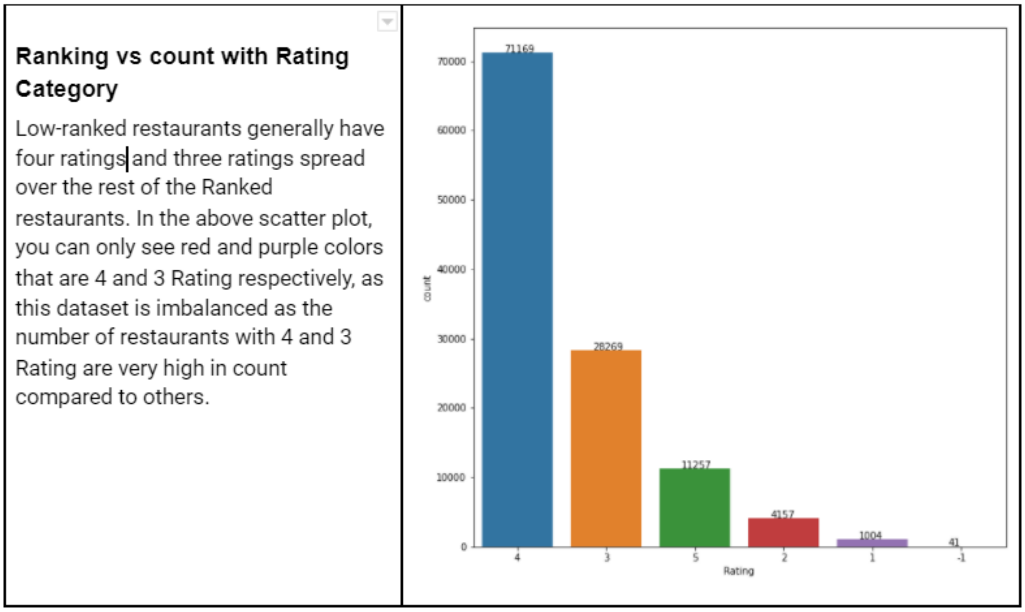
- **Removing Rows** - After removing the above columns now left with seven columns, remove all those rows with null values in almost all six columns that columns fetched in the Exploration stage. (121061, 7)

Rounding



Ranking vs Number of Reviews with Rating Category

Round off the Rating column so that majorly left with five categories



Ranking vs count with Rating Category

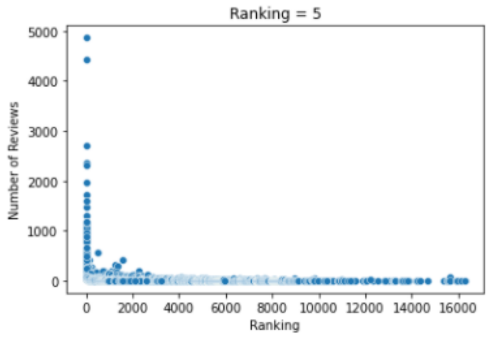
Low-ranked restaurants generally have four ratings and three ratings spread over the rest of the Ranked restaurants. In the above scatter plot, you can only see red and purple colors that are 4 and 3 Rating respectively, as this dataset is imbalanced as the number of restaurants with 4 and 3 Rating are very high in count compared to others.

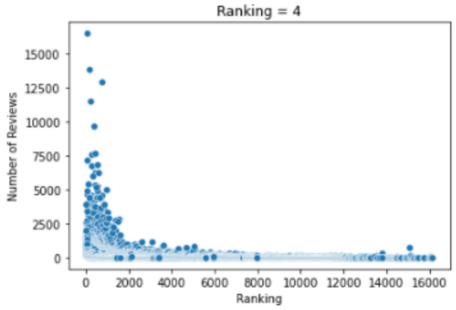
Scatter Plot for each Rating

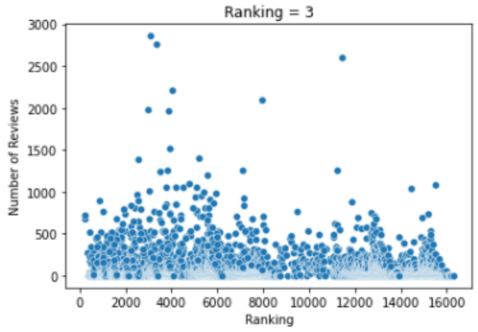
Five and Four Rating is received mostly by low ranked restaurants, whereas three Rating is the most versatile one spread over all the ranks.

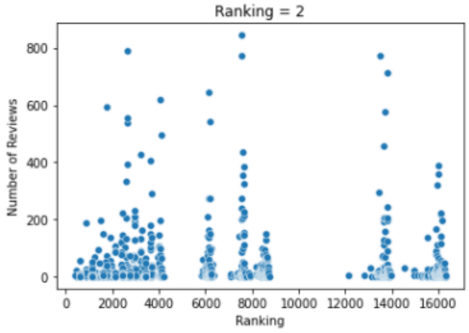
Remove rows with Rating = -1, As they don't have their respective Reviews.

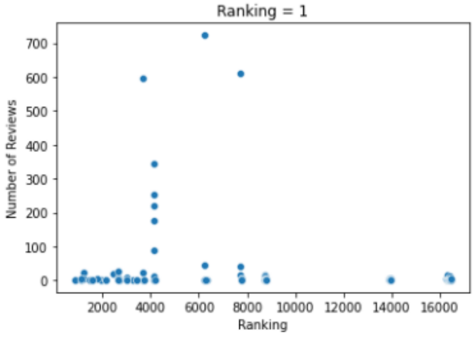
New Data Shape-(115856, 7)











Chapter 3

Methodology

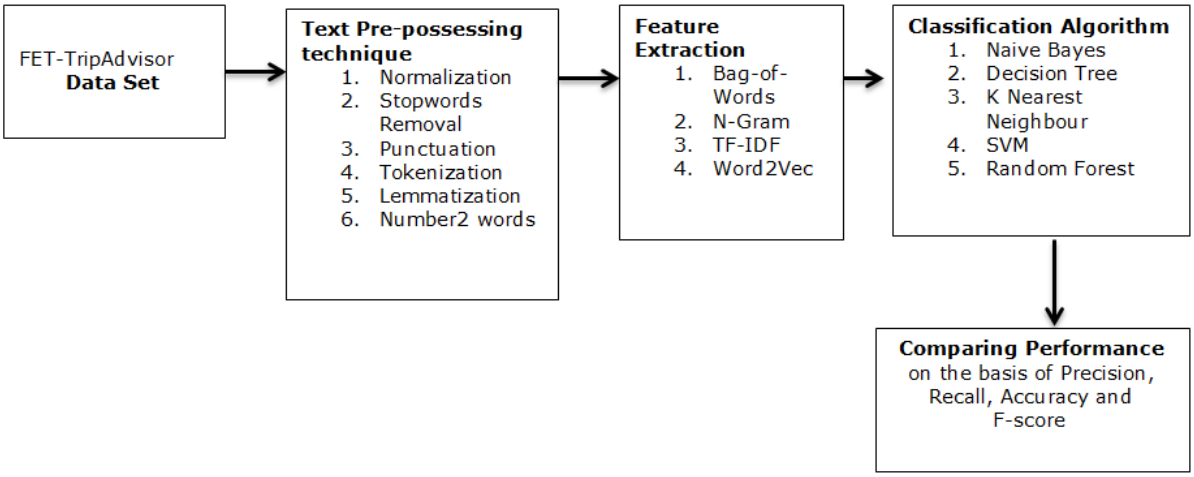


Figure 3.1: Methodology

3.1 Text Pre-processing

In this section are described the preprocessing steps applied to make the data more suitable for data mining. Different strategies have been adopted to turn raw input into essential data.

Following is listed a set of operations that have carried out in the text cleaning process(pre processing):

- **Normalization** - Originally Data was in the string format which is converted into list of sentences(strings).

- **Stopwords Removal** - In the experiments, a stopwords list was implemented, excluding some words (like the' , 'a', 'is'), due to being 'fairly good' indicators of negative class as suggested from the data.

- **Punctuations** - Removing punctuation marks and special characters.

- **Tokenization** - Before classification experiments, tokens (i.e. words) in the dataset were pre-processed using a Stemmer. For this purpose, an english Stemmer from nltk was implemented.

- **Lemmatization** - It considers the context and converts the word to its meaningful base form

- **Number 2 words** - Convert all digits(or numbers) to word format and remove.

3.2 Word Clouds after pre-processing

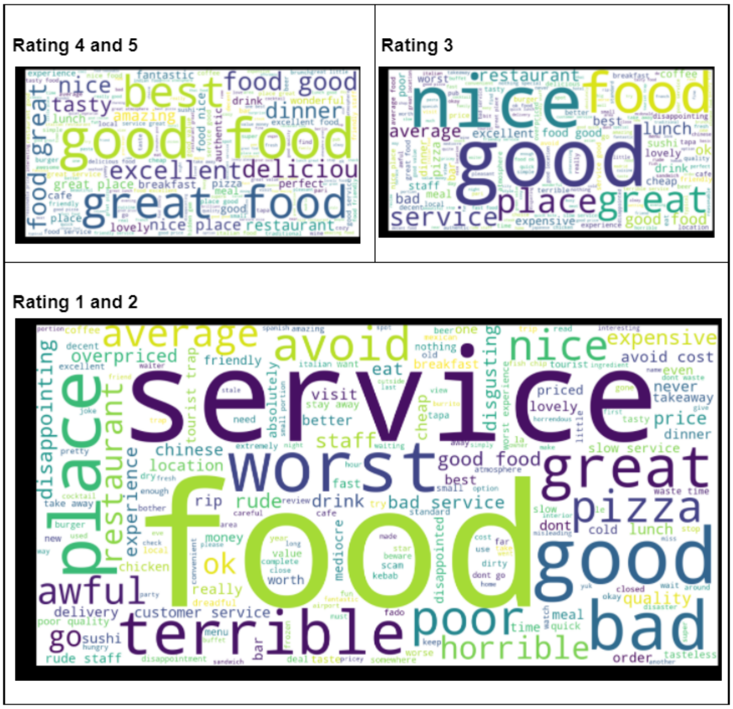


Figure 3.2: Ratings

3.3 Feature Extraction

Bag Of Words

It is a method to extract features from text documents. Machine Learning algorithms can use these extracted features for training machine learning algorithms. It generates the vocabulary of all the unique words occurring in all the documents in the training data.
So, BOW is a collection of words to represent a sentence with word count while disregarding the order in which they appear.
BOW is an approach widely used with

- Natural language processing
- Information retrieval from documents
- Document classifications

N-gram

It is the sequence of N words, a 2-gram (or bigram) is a two-word sequence of words like “you do”, “do this”, or ”this task”, and a 3-gram (or trigram) is a three-word sequence of words like “you do this”, or “do this task”.N-gram models are used in probability, communication theory, computational. linguistics, computational biology and data compression.

Benefits of these models are simplicity, scalability with larger n, a model thus stores more context with space–time tradeoff.

TF-IDF

TFIDF is a way to convert textual data to numeric form, and means Term Frequency-Inverse Document Frequency. The numeric vector it yields is the product of these two terms; TF and IDF.
Relative term frequency is calculated for each term within each document as below.

$$tf_{t,d} = \frac{n_{t,d}}{Number\ of\ terms\ in\ the\ document}$$

$$idf_t = \log \frac{number\ of\ documents}{number\ of\ documents\ with\ term\ 't'}$$

$$(tf\_idf)_{t,d} = tf_{t,d} * idf_t$$

Word2Vec

Word2vec is a method to create word embeddings efficiently and with pre-trained models. The benefit of using this is that it establishes word embeddings such that similar meaning words are near to each other in the vector space. And to represent the whole sentence in the vector form, we used an approach of averaging out all the vector representations of the words in a sentence.

3.4 Classification Algorithms

Naïve Bayes

This algorithm is based on Bayes' theorem with the assumption of independence between features. These classifiers work well in real-world situations like document classification and spam filtering.It needs a small amount of training data to estimate, is extremely fast compared to more sophisticated methods.

Decision Tree

Provided a dataset of attributes along with its classes, a decision tree generates a sequence of rules that are used to classify the dataset.It is simple to visualise/understand, can handle both categorical and numerical data and requires little data preparation.

K Nearest Neighbor

This algorithm is a supervised classification/regression algorithm. To label a new point, it looks at the nearest neighbors to that new point, and has those neighbors vote. So whichever label, the most of its nearest neighbors have is the label for the new point. Also, here "k" represented in K-Nearest Neighbors is the number of nearest neighbors it looks at. It is supervised because we are trying to classify a point based on the known classification of other points.

Support Vector Machine

It is a representation of the training set as points in space which are separated into various categories by a clearly visible gap(hyperplane) which is as wide as possible. Test points are then mapped into this space and predicted for a category based on where they fall.

Effective in high dimensional spaces, memory efficient as it uses a subset of training points in the decision function.

Random Forest

It's a meta-estimator which fits decision trees on sub-samples of datasets and uses average to improvise the predictive accuracy of the model and is very effective in minimizing the over-fitting. This classifier is more accurate than decision trees in many cases.

Chapter 4

Results

When to use the macro-average method ?

Macro averaging reduces the multiclass predictions down to multiple sets of binary predictions, calculates the corresponding metric for each binary case, and averages the results together. Example, If there were levels A, B, and C, macro averaging reduces the problem to multiple one-vs-all comparisons in the multiclass case. The truth and estimate columns are recorded such that the only two levels are A and other, and then precision is calculated based on those recorded columns, with A being the "relevant" column. This process is repeated to get a total of 4 precision values.

Why is accuracy not always a good metric?

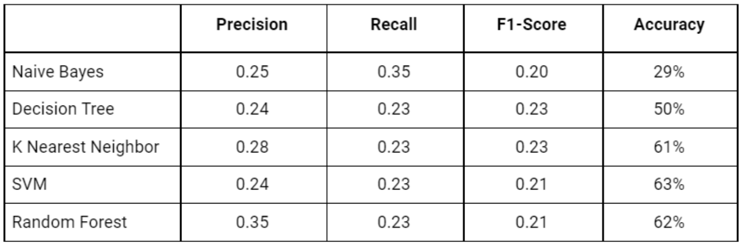
Accuracy is the most widely used metric for evaluating classification models. It is easy to calculate, interpret and is a single number to tell the model's capability. When the skew in the class distributions is severe, accuracy can become unreliable.

Example, if out of 100 data items 90 belongs to first class and rest of the 10 data items from the second class. If the model classifies correctly only 80 labels of first class. Accuracy of the model is 80 percent although we know that model is very bad at predicting class labels of second class.

The following table shows the macro-averages of the metrics(Precision,Recall, F1-Score) obtained on the development dataset:

Note- For k-nearest neighbors k=5 is used. Accuracy is in Percentage (

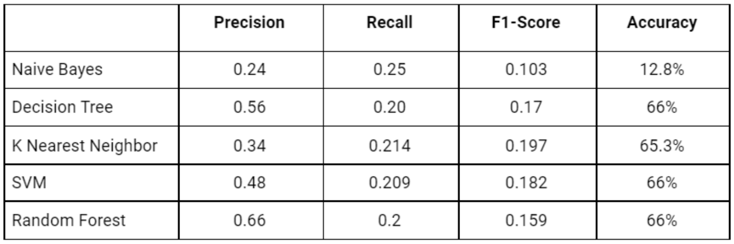
4.1 Bag of Words Result



	Precision	Recall	F1-Score	Accuracy
Naive Bayes	0.25	0.35	0.20	29%
Decision Tree	0.24	0.23	0.23	50%
K Nearest Neighbor	0.28	0.23	0.23	61%
SVM	0.24	0.23	0.21	63%
Random Forest	0.35	0.23	0.21	62%

Figure 4.1: Bag of Words

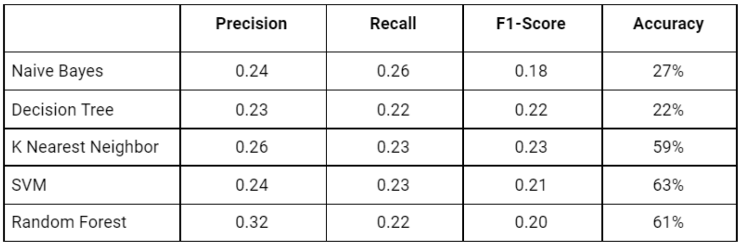
4.2 N-gram Result



	Precision	Recall	F1-Score	Accuracy
Naive Bayes	0.24	0.25	0.103	12.8%
Decision Tree	0.56	0.20	0.17	66%
K Nearest Neighbor	0.34	0.214	0.197	65.3%
SVM	0.48	0.209	0.182	66%
Random Forest	0.66	0.2	0.159	66%

Figure 4.2: N-gram

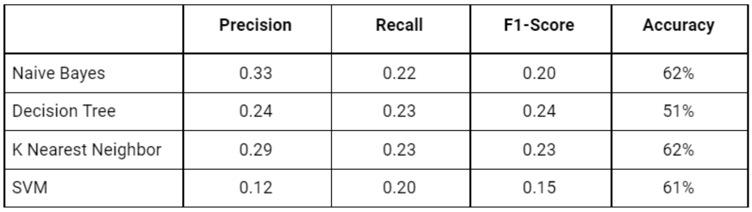
4.3 TF-IDF Result



	Precision	Recall	F1-Score	Accuracy
Naive Bayes	0.24	0.26	0.18	27%
Decision Tree	0.23	0.22	0.22	22%
K Nearest Neighbor	0.26	0.23	0.23	59%
SVM	0.24	0.23	0.21	63%
Random Forest	0.32	0.22	0.20	61%

Figure 4.3: TF-IDF

4.4 Word2Vec Result



	Precision	Recall	F1-Score	Accuracy
Naive Bayes	0.33	0.22	0.20	62%
Decision Tree	0.24	0.23	0.24	51%
K Nearest Neighbor	0.29	0.23	0.23	62%
SVM	0.12	0.20	0.15	61%

Figure 4.4: Word2Vec

4.5 Metric Visualization

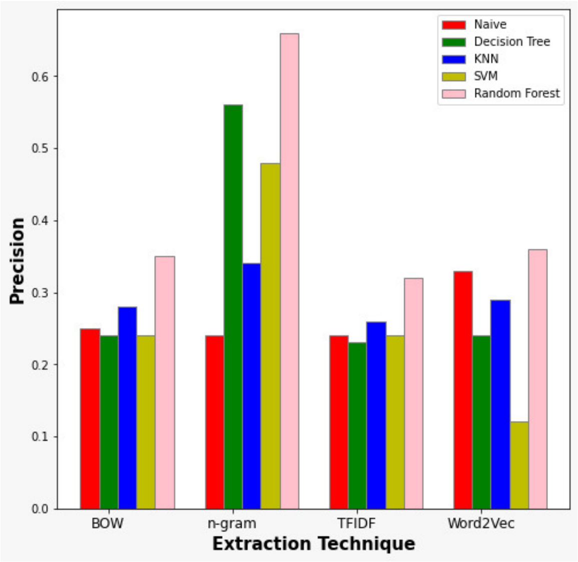


Figure 4.5: Precision Score

N-Gram precision score is more compare to other

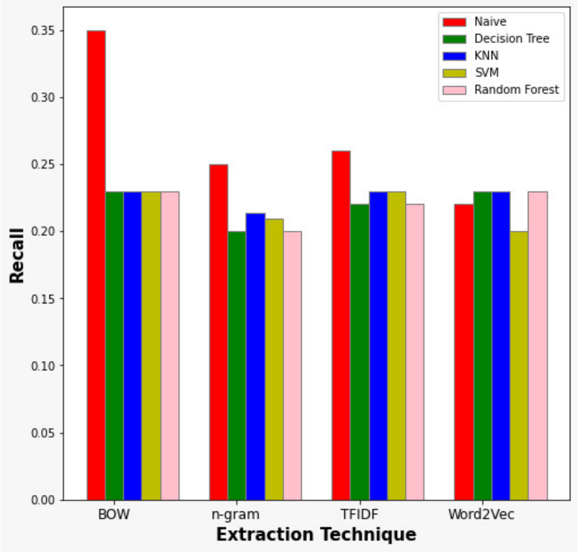


Figure 4.6: Recall

BOW Recall score is more compare to other

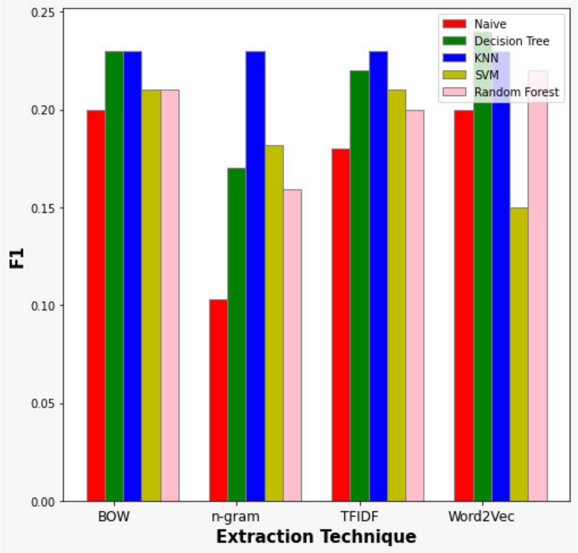


Figure 4.7: F1 Score

Word2Vec F1 score is more compare to other

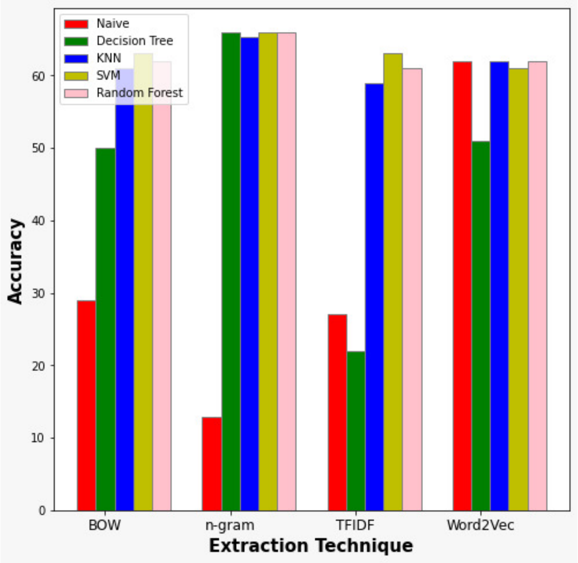
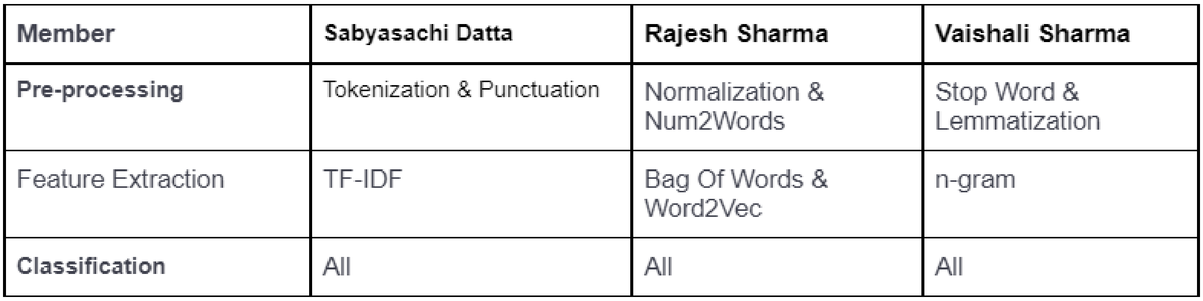


Figure 4.8: Accuracy

N-Gram Accuracy is more compare to other

Chapter 5

Individual tasks



Member	Sabyasachi Datta	Rajesh Sharma	Vaishali Sharma
Pre-processing	Tokenization & Punctuation	Normalization & Num2Words	Stop Word & Lemmatization
Feature Extraction	TF-IDF	Bag Of Words & Word2Vec	n-gram
Classification	All	All	All

Note: Overall understanding of the project- by all three group members

Bibliography

[1] Research Paper

https://www.researchgate.net/publication/333524196_The_Impact_of_Features_Extraction_on_the_Sentiment_Analysis

[2] Dataset of TripAdvisor

https://www.kaggle.com/damienbeneschi/krakow-ta-restaurans-data-raw

[3] Refer scikit learn

https://scikit-learn.org/stable/supervised_learning.html