THESIS PAPER

Bangla Fake Food Review Detection Using Machine Learning Algorithms

Prepared by
Name: Md Ashraf Uddin
Student ID: 2018-2-58-043

Name: Tusher Paul
Student ID: 2018-3-50-016

Name: Md Riaz Ahmmed
Student ID: 2018-2-50-022

Supervised by
Dr. Mohammad Arifuzzaman
Associate Professor

This Thesis Paper is Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Information and Communications Engineering

DEPARTMENT OF ELECTRONICS & COMMUNICATIONS ENGINEERING

EAST WEST UNIVERSITY

APPROVAL

The thesis paper titled “Bangla Fake Food Review Detection Using Machine Learning Algorithms” submitted by Md Ashraf Uddin(Student ID:2018-2-58-043),Tusher Paul(Student ID: 2018-3-50-016) and Md Riaz Ahmmed (Student ID: 2018-2-50-022) to the Department of Electronics and Communications Engineering, East West University, Dhaka, Bangladesh has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Information and Communications Engineering and approved as to its style and contents.

Approved By

- -

(Supervisor)

Dr. Mohammad Arifuzzaman

Associate Professor

ECE Department

East West University

Dhaka, Bangladesh

DECLERATION

We certify that neither this University nor any other University has previously accepted our work for the purpose of awarding a degree. To the best of my knowledge and belief, this essay does not use any content that has already been published or authored by another individual unless it is specifically cited in the text. We thus state that the research we conducted under the direction of Dr. Mohammad Arifuzzaman, Associate Professor, Department of Electronics & Communications Engineering, East West University, Dhaka, Bangladesh, resulted in the work described in this thesis paper.

Countersigned

(Supervisor)

Dr. Mohammad

Arifuzzaman

Signature

Md Ashraf Uddin

Student ID: 2018-2-58-043

Signature

Tusher paul

Student ID: 2018-3-50-016

Signature

Md Riaz Ahmed

Student ID: 2018-2-50-022

DEDICATION

This paper is

dedicated To

Our beloved parents and honorable teachers

ACKNOWLEDGEMENT

We would like to extend our gratitude to our supervisor, Dr. Mohammad Arifuzzaman for his inspiring words and sincere guidance to complete this work in an efficient way. His constant support gave us the assurance we needed to work on this paper. Our completion of this paper could not have been accomplished without the support of our honorable supervisor. We are truly honored and thankful to have him as a supervisor in the journey of our thesis. We are also grateful to our honorable faculty members and office stuffs who rendered their help during the period of our thesis work.

ABSTRACT

In recent years, Fake Review Detection has emerged as a key and well-liked topic for both business and research due to the explosive growth of e-commerce platforms. In today's e-commerce, internet reviews are crucial for helping consumers make the best decisions. Numerous social media platforms and marketing websites host millions of reviews of various goods and services. Customers post reviews of products based on their personal experiences, and other buyers can learn more from these reviews before making a purchase choice. However, since there are no restrictions on what can be posted in a review on any internet platform, it can often be quite difficult to identify the real reviews. According to them, anyone may submit reviews, which increases the quantity of bogus reviews and might provide false information, misleading a buyer. These reviews, whether they are phony or real, have a big impact on an organization's revenue and reputation. Positive reviews typically increase sales and profit while negative reviews negatively impact a company's reputation. We are motivated by this situation to discern between bogus and real reviews. In this study, we cover several machine learning techniques (Naive Bayes, Random Forrest, K-Nearest Neighbor, Logistic Regression, Support Vector Machine (Linear), Decision Tree), and we try to determine whether the review is credible or not using those algorithms. We used a confusion matrix and examined the outcomes of the experiment. We also talk about some of the difficulties we had writing this thesis and some of our future plans for it.

Table of Contents

Chapter:1(Introduction)

1.1: Introduction

Sentiment analysis is the extraction and comprehension of the sentiments defined in a text document. The avalanche of information in the numerous social media platforms such as Twitter, Facebook, and LinkedIn have provided consumers with new avenues for expressing their opinions on specific products, people, and locations. User feedback always takes the form of textual content. Every day, millions of text messages are transmitted over social media or e-commerce websites. Investigating and assessing the opinion's sentiment is a crucial duty to do. NLP with artificial intelligence capabilities and text analytics are used to evaluate whether an opinion's sentiment is good, negative, or neutral. Opinion mining and sentiment analysis are independent of any specific domain or platform. It permeates all social media networks, healthcare, management, the economics, and many others, and is also highly beneficial to the expansion of a great number of businesses and organizations [1]. Reviews are statements made by consumers to other consumers in which they offer advice, opinion, or experience in regard to a product available on the market. Online reviews are very prevalent in current internet era. People are very eager to discuss their experiences with utilizing various products. They publish their opinion or review on several online portals. These reviews are highly beneficial for various businesses or individuals that are interested in purchasing comparable goods or services. Based on such reviews, individuals can form an opinion and decide before choosing. The number of customer reviews has rapidly expanded in recent years, and these reviews have a big impact on the decision of the buyer. In other words, consumers decide whether to buy a product after reading reviews on social media or to change their minds. Consumer reviews therefore provide an important service to people [2]. Reviews might be favorable or unfavorable. Negative reviews hurt the market or any firm whereas positive reviews always result in enormous financial gains. Both of these reviews, whether favorable or unfavorable, have a significant impact on customers' decisions because they always check customer reviews before deciding whether to purchase a particular product or use a particular service from a company or restaurant, and they base their decisions on those reviews. Because of this, consumer reviews or comments are developing into a very valuable source of knowledge for making any selection for the customers. For instance, people always look up a restaurant's design or cuisine rating before going there. They navigate to various restaurants and menus on their pages and read evaluations or client testimonials about the establishments and their offerings. They decide whether to visit a particular restaurant based on the favorable or negative ratings they have received. As a result, historical analyses rose to the top of many web services' lists of sources of information that users can trust [3]. Since people are free to express or share their opinions on various social media platforms, the websites that feature customer reviews have a significant impact on markets. People are free to publicly share their opinions about any business at any time, with no restrictions or duties. Due to the lack of restrictions, many of these businesses resort to using dishonest techniques of all kinds. Numerous businesses occasionally unfairly advertise their products or services on various social media platforms, and they also unfairly criticize their rivals, misleading the public. The owner of the business will occasionally engage individuals to post unfavorable reviews about their competitors' items in order to promote the business and increase profits. Reviews are regarded as real feedback regarding good or bad services, therefore any attempt to skew them by including false or misleading information is seen as dishonest behavior and is accompanied by the label "fake reviews" [3]. A spammer is someone who publishes false customer reviews. Such negative reviews make us question the validity of the comments and reviews that are now available on various social media platforms. The consumer finds it difficult to distinguish between bogus and real evaluations, and we begin to wonder whether all of the submitted reviews are genuine or not. In such a circumstance, it is imperative that we ascertain the veracity of the customer's feedback or opinion; hence, the field of research known as "Detection of Fake Reviews" has become and continues to be one of the most fruitful and challenging. Numerous studies and investigations have been conducted on the detection of fake and genuine reviews, as well as the difficulties involved. Initial identification of fake reviews was introduced by Jinal et al [4]. There are several techniques for spotting false reviews. Two of these are deep learning and machine learning approaches. The classification of fake and non-fake reviews is the primary function of fake review detection. In this study, we applied certain Machine Learning Algorithms to try and distinguish between fake and real reviews. Here, we attempted to compare these algorithms in terms of calculating the accuracy for our dataset. The detection of false reviews is assisted by the use of machine learning techniques, making it easier for researchers interested in this area to select the most effective machine learning approach. Additionally, this paper will make it simple for the reader to understand the field of Bangla fake review identification and its significance.

1.2 Problem Statement

Fake reviews don't match up with honest opinions of goods or services. Therefore, fake reviews are untrue, unreliable, and dishonest reviews that are written by a variety of users, including online shoppers and retailers. Either the polarity of fake reviews is positive or negative. Reviews that contain positive statements about a product are of the positive polarity, while those that contain negative statements about a product are considered to be of the negative polarity. The perception of the product is significantly impacted by each review. Reviews are essential for establishing and preserving a solid reputation on the e-commerce marketplace. In addition, these are useful in helping end users decide whether to buy any products. These false reviews hurt numerous businesses that sell high-quality products since they not only confuse buyers when making decisions about what to buy.

The necessity to identify fraudulent reviews is growing, making fake review detection an appropriate study issue nowadays.

Many studies have been conducted in this area to determine the best method for identifying bogus reviews. In our example, we attempted to simulate our research by using a Bangla dataset and a machine learning algorithm. Our ultimate objective for this study in Bangla fake review detection can be summed up as follows:

i. Choosing the precise Bangla dataset that is pertinent to this study issue and pre-processing the data before using various methodologies.
ii. Using machine learning algorithms to analyze datasets
iii. Displaying a comparison of the outcomes of machine learning approaches
iv. Assessing the outcomes and drawing conclusions from this study's findings

1.3 Motivation

Buyers are driven to internet marketplaces in addition to conventional markets and superstores because they can avoid heavy traffic, lengthy waits, crowded settings. Customers increasingly rely on product evaluations for information before making any purchases due to the rapid advancement of technology. On the other hand, fake reviews reduce the value of online reviews by creating a false impression of the caliber of the product. Therefore, fraudulent review detection has become quite important in recent times. Unfortunately, automatic detection hasn't been very effective at this challenging task so far. More than 80% of American buyers say they consult online reviews before making a purchase (Smith and Anderson, 2016). Despite being the oldest of Generation Z are still young adults, yet they have a substantial spending power. Marketers must comprehend the specific needs and spending preferences of this generation in order to be relevant [by Coral Ouellette on January 7, 2022]. In terms of purchasing power, the generation born after 1998 is predicted to have $44 billion. Now, 93% of parents say that their Gen Z child affects the family budget. This generation will account for 40% of all consumer purchases in a few short years [by Coral Ouellette on January 7, 2022]. 95% of the people in this group use their smartphones for at least 10 hours a day. They therefore have a 2X higher likelihood of shopping on mobile devices than millennials do. However, 85% of them use social media platforms [by Coral Ouellette on January 7, 2022] to learn about new products. We can predict the future of internet buying by keeping in mind these figures. Evaluations are one of the most important things that consumers consider when making purchases, therefore dishonest actors may be persuaded to employ or utilize automated methods to create fake reviews to increase the allure of their goods and services or damage the reputations of rivals. We may anticipate that a significant portion of consumers will switch to online shopping soon, which will result in a rise in the number of evaluations on various online platforms. There is a possibility that the quantity of phony reviews may rise since individuals are more inclined to rely on those reviews to gauge the quality of products and base their purchasing decisions thereon. Therefore, obtaining an online platform for detecting fake reviews is a very vital tool for the present and the future. Fake review detection is still a very effective and demanding study subject. We decide to work in this area because it is crucial to distinguish between bogus and real reviews.

Bangla dataset- Bangla fake food reviews—were used in this study. On this Bangla fake food review dataset, we applied machine learning technique. Utilizing this algorithm on our targeted data is motivated by the need to compare the results and determine which algorithms are most effective at identifying false reviews. The results of this research will be used to construct a platform in the future that detects bogus reviews automatically.

1.4 Thesis Organization:

There are a total of seven chapters in our thesis. Introduction, problem statement, motivation, and thesis arrangement are all included in the first chapter of our study. We gave the literature study in the second chapter, where we briefly discussed related to our work has been the subject of prior works. We covered the implemented machine learning algorithms in our third chapter. We introduced the fundamental concept of bogus review identification in the fourth chapter. Our research technique is provided in the fifth chapter. Our research findings and analyses using a machine learning method are presented in the sixth chapter. The final words addressing our entire thesis effort are provided in chapter seven.

Chapter:2(Literature Review)

In recent years, both businesses and scholars have focused a lot of effort on the identification of fake reviews. In order for reviews to reflect actual user experiences and opinions, fake reviews must be recognized [5].

Numerous research papers, essays, and survey papers have been published; these are the works that are immediately and most pertinently related. The following are some of those from which the present study took its cues:

In author [6] used sentiment analysis to identify and categorize the positive and negative sentiment expressed in a piece of text where consumers can submit reviews with a specified rating on e-commerce websites like Amazon.com. In the paper, they have sought to construct sentiment analysis related to product ratings and text reviews.

In author [7] used Sensory analysis which has gained a wide interest in natural language processing (NLP). Their paper focuses on SA in the context of Bangla language. A novel rule-based algorithm termed as Bangla Text Sentiment Score (BTSC) is developed for detecting sentence sentiment.

In author [8] analyzed data from social networking sites, where a business entity can be benefited in their product marketing. They have developed a natural language processing (NLP) based pre-processed data framework to analyze sentiment.

In author [9] analyzed the Movie reviews using various techniques like Naïve Bayes, K-Nearest Neighbor and Random Forest. Social media and other online platforms contain a huge amount of the data in the form of tweets, blogs, and updates.

In author [10] took dataset has taken from Amazon which contains reviews of Camera, Laptops, Mobile phones, tablets, TVs, video surveillance. They applied machine learning algorithms to classify reviews that are positive or negative. Naïve Bayes got accuracy 98.17% and. Support Vector machine got accuracy 93.54% for Camera Reviews.

In author [11] presented a new feature for classifying the tweets as positive, negative and extract people's opinion about products. Using Machine Learning approach, it is possible to identify the effect of domain information in sentiment classification.

In author [12] purposed to find the sentiment from a Bengali paragraph which is happy or sad. After preprocessing the data, they applied six algorithms to predict almost high accuracy. The maximum accuracy which is 86.67%.

In author [13] used sentiment analysis, a given text has been categorized into several emotions. Their paper deals with six individual emotion classes-happy, sad, tender, excited, angry and scared.

In author [14] Make a system which analyze customer's feedback and provide a ratio of positive and negative feedback written in Bangla.

In author [15] proposed two neural network models that integrate traditional bag-of-words as well as the word context and consumer emotions. The proposed systems perform well on all datasets, irrespective of their sentiment polarity and product category.

In author [16] proposes a feature framework for analyzing fake reviews in the consumer electronics domain. They have reached an 82% F-Score on the classification task and the Ada Boost classifier has been proven to be the best by statistical means.

In author [17] aimed to build a machine-learning model to analyze the sentiment of the reviews. They have collected more than one thousand food reviews from various online platforms like Foodbank, Food panda, Hungry naki, Shohoz food, Pathao food, etc., and labeled them.

In author [18] proposes DL models for SA on Bangla text using an extended lexicon data dictionary. Preprocessed texts are formatted as a vectorization of words of unique numbers of pre-trained word embedding models.

Chapter:3(Machine Learning Algorithms)

3.1Machine Learning Algorithms

Humans learn from past experiences. Humans store their past experiences and when it comes to the action, humans use their experiences and solve those problems in real time. On the other hand, the machine follows a set of instructions that are given by humans. The machine is as straight as a fellow instructor. But machine learning is quite different. A machine learning algorithm is a computational process that uses input data to achieve a desired task without being literally programmed (i.e., “hard coded”) to produce a particular outcome. These algorithms are "soft programmed" in the sense that they automatically adjust or adapt their design as a result of repetition (i.e., experience) to get better and better at doing the target objective. What if humans can train the machines. Human set the previous data on machines and machines can act in a faster way. Basically, that is called machine learning. Machine learning is an essential branch of AI.

Machine Learning is in the application of artificial intelligence (AI) that provides systems the ability to enables computers to grow, modify and learn by themselves when they are exposed to a new data. Machine learning's objective is to mimic how humans (and other sentient organisms) learn to process sensory (input) data in order to achieve a goal. This objective could be a pattern recognition challenge in which the learner must discriminate between apples and oranges. Although each apple and orange are distinct, we can typically distinguish one from the other. Cybercrime is growing rapidly as there are more people using the internet. Phishing attacks using phishing URLs, phishing emails, and phishing websites are fairly common. Both the attackers and the users who seek to stop this attack are entering a new era thanks to machine learning [19].

Machine learning refers to the field of study, which enables machines to keep improving their performance without the need for programming. Nowadays Machine learning is the process of predicting and analyzing Breast Cancer. From data, helpful information. The scope of these approaches' applications and potentials can be found in the medical data.

The prediction section is typically made simpler by classification approaches. Generally speaking, a doctor must spend a lot of time and effort classifying breast cancer. Therefore, it is imperative to identify cancer using various automatic diagnostic methods [20].

Through machine learning, your software and bots can learn new things always and give better results [20].

Types of Machine Learning:

There are mostly common three types of machine learning and those are,

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

3.1.1: Supervised Learning

Supervised learning, also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. Supervised learning is when you provide the machine with a lot of training data to perform a specific task.

In Supervised learning, you train the machine using data that is well "labeled." It means some data is already tagged with correct answers. It can be compared to learning in the presence of a supervisor or a teacher. It is the most accessible type of Machine Learning to implement, and it's also the most common one. Supervised machine learning helps you to solve various types of real-world computation problems [21].

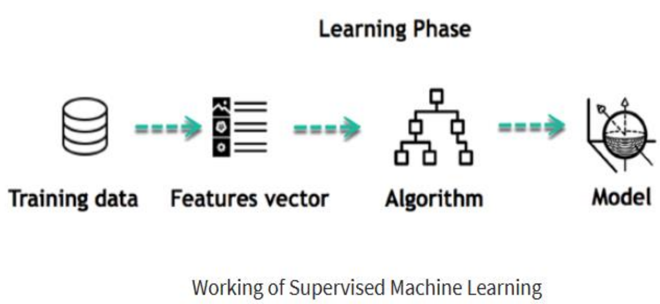

Figure 3.1: Supervised Learning

[Source Link: https://www.guru99.com/supervised-machine-learning.html#:~:text=Supervised%20Machine%20Learning%20is%20an,already%20tagged%20with%20correct%20answers]

Naive Bayes Classifier Algorithm

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Naive: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes' Theorem:

Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability. The Bayes theorem is used to create the Nave Bayes classifier, which is based on the premise of strong independence [22].

The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(A) \times P(B \vee A)}{P(B)} \quad \ldots\ldots\ldots.[1]$$

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

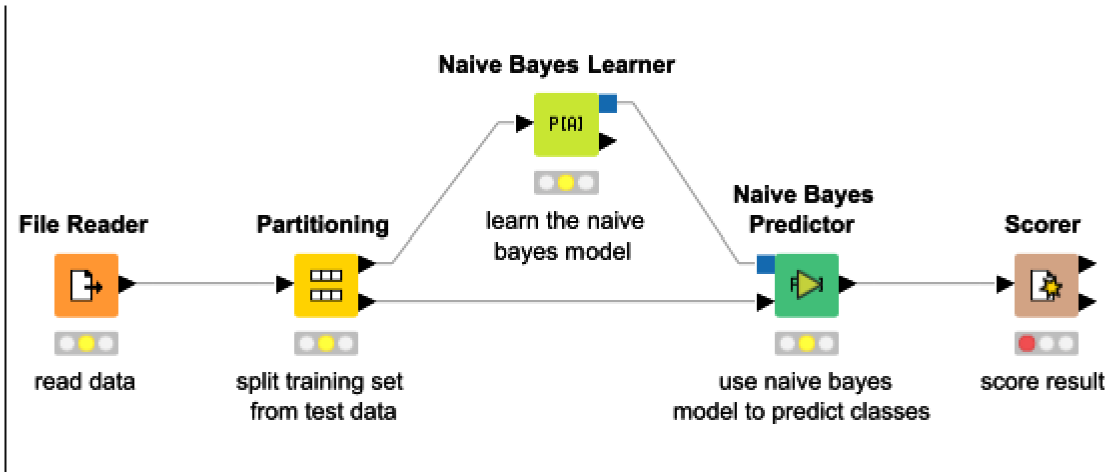

Figure 3.2: How Naïve bayes Algorithm works
[Source Link: https://i.stack.imgur.com/0QOII.png]

Logistic Regression in Machine Learning

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables [23].

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

A Logistic Regression model is like a Linear Regression model, except that it uses a more sophisticated cost function, which is known as the 'Sigmoid function' or the 'logistic function' instead of a linear function.

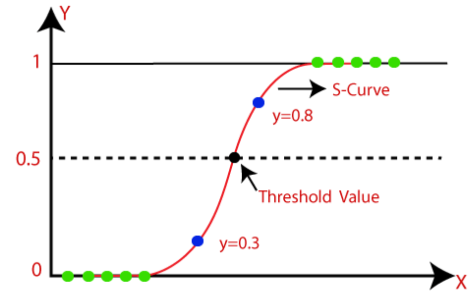

Figure 3.3: Logistic function models the conditional probability of the response.
[Source Link: https://static.javatpoint.com/tutorial/machine-learning/images/logistic-regression-in-machine-learning.png]

The Logistic Regression can be obtained from the Linear regression equation. The mathematical step to get Logistic Regression equation are given below:

- The equation of the straight-line cane be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n \qquad \text{.........[2]}$$

- Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}; \text{ 0 for y= 0, and infinity for y=1} \qquad \text{..............[3]}$$

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n \qquad \text{............[4]}$$

Logistic Regression can be classified into three types:

Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.

Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"

Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Support Vector Machine Algorithm

Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

Each data item is plotted as a point in n-dimensional space (where n is the number of features you have), with the value of each feature being the value of a certain coordinate in the SVM algorithm. Then we accomplish classification by locating the hyper-plane that clearly distinguishes the two classes.

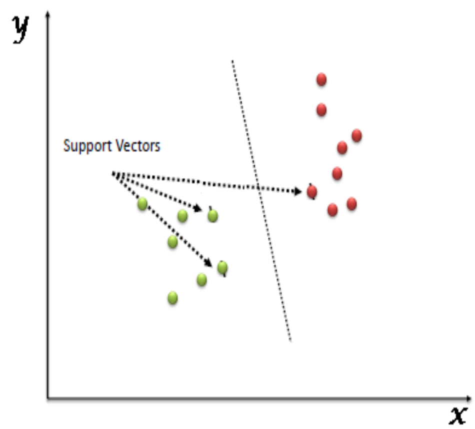

Figure 3.4: Support Vector
https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_21.png

Support Vectors are simply the coordinates of individual observation. The SVM classifier is a frontier that best segregates the two classes (hyper-plane/ line) [24].

Decision Tree

The most powerful and widely used tool for categorization and prediction is the decision tree. Each internal node symbolizes a test on an attribute, each branch represents a test outcome, and each leaf node (terminal node) stores a class label.

By separating the source set into subgroups based on attribute value tests, a tree can be "trained." Recursive partitioning is the process of repeating this method on each derived subset. When all of the subsets at a node have the same value of the target variable, or when splitting no longer adds value to the predictions, the recursion is complete. No domain expertise or programming skills are required to build a decision tree classifier.

Instances are classified using decision trees by sorting them down the tree from the root to a leaf node, which provides the classification. As indicated in the above diagram, an instance is classified by starting at the root node of the tree, checking the attribute specified by this node, and then progressing along the tree branch according to the attribute value. This procedure is then repeated for the new node's subtree [25].

Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems. Random Forest is a classifier that consists of a few decision trees on various subsets of a given dataset and takes the average to improve the dataset's projected accuracy. A forest is a grouping of trees, while a random forest is a grouping of categorization trees. The creation of a tree in which the members of the class variable dwell on the leaf nodes and the entities of other dependent variables reside on the intermediate nodes is known as a classification tree. Random Forest is capable of performing both Classification and Regression tasks [26].

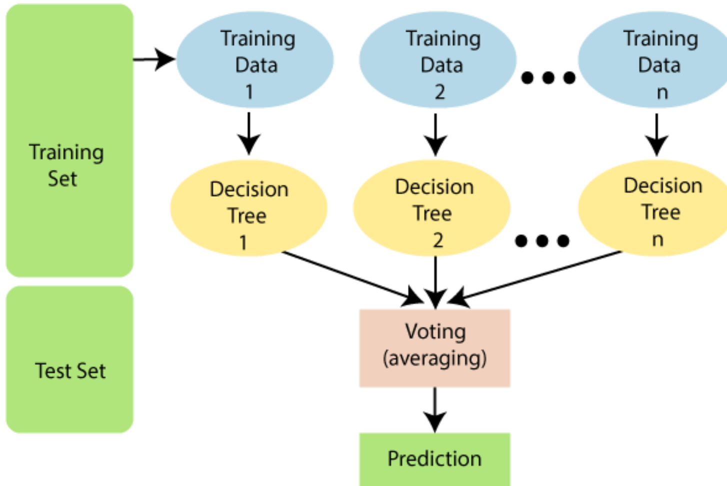

Figure 3.5: Random Forest
[Source Link: https://static.javatpoint.com/tutorial/machine-learning/images/random-forest-algorithm.png]

Assumptions for Random Forest

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Example: Suppose there is a dataset that contains multiple fruit images. So, this dataset is given to the Random Forest classifier. The dataset is divided into subsets and given to each decision tree. During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random Forest classifier predicts the final decision. Consider the below image:

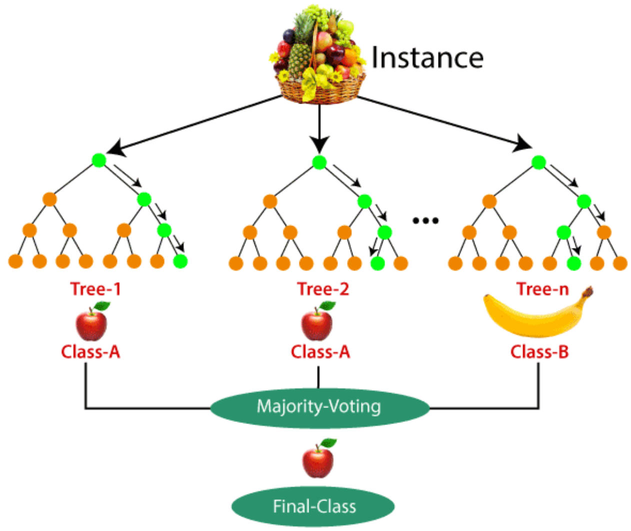

Figure 3.6: Example of Random Forest classification
[Source Link: https://static.javatpoint.com/tutorial/machine-learning/images/random-forest-algorithm2.png]

3.1.2: Unsupervised Learning:

Unsupervised learning is a type of machine learning algorithm that looks for patterns in a dataset without pre-existing labels. As the name suggests, this type of machine learning is unsupervised and requires little human supervision and prep work. Because unsupervised learning does not rely on labels to identify patterns, the insights tend to be less biased than other forms of AI. **Unsupervised Learning Algorithms** allow users to perform more complex processing tasks compared to supervised learning. Although, unsupervised learning can be more unpredictable compared with other natural learning methods [27].

Unsupervised learning models are used in the following ways:

- **Clustering:** This is the process of finding similarities among unlabeled data and grouping them together.
- **Association:** This unsupervised learning method finds relationships between the data in a given dataset.

- **Dimensionality Reduction:** This machine learning technique is used when the number of features in a dataset is too high. This technique reduces the number of inputs into a more manageable size all while preserving the data integrity.

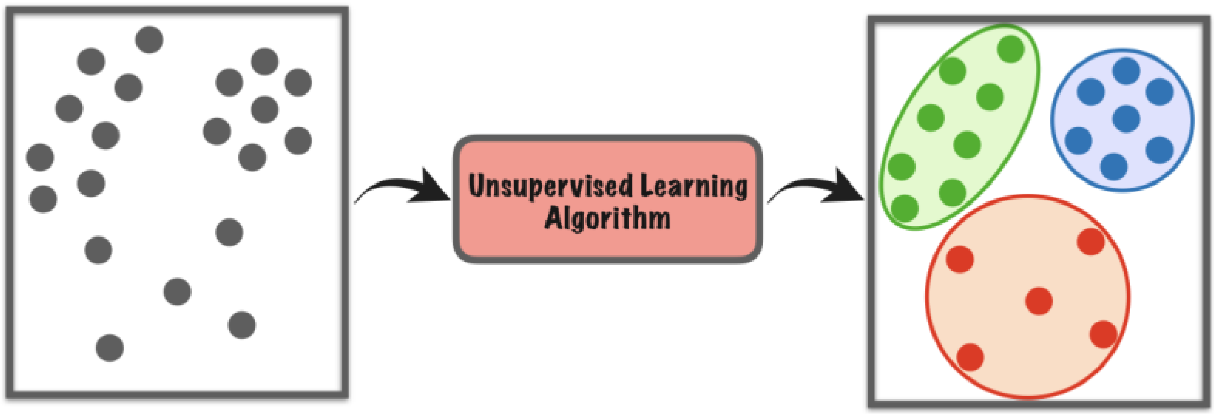

Figure 3.7: Unsupervised Learning
[Source Link: https://unsupervised.com/resources/blogs/what-is-unsupervised-learning/]

3.1.3: Reinforcement Learning:

Reinforcement learning is the training of machine learning models to make a sequence of decisions. The agent learns to achieve a goal in an uncertain, potentially complex environment. In reinforcement learning, an artificial intelligence faces a game-like situation. The computer employs trial and error to come up with a solution to the problem. To get the machine to do what the programmer wants, the artificial intelligence gets either rewards or penalties for the actions it performs. Its goal is to maximize the total reward [28].

Reinforcement learning (RL) is learning what to do, given a situation and a set of possible actions to choose from, in order to maximize a reward. The learner, which we will call agent, is not told what to do, he must discover this by himself through interacting with the environment. The goal is to choose its actions in such a way that the cumulative reward is maximized. So, choosing the best reward now, might not be the best decision, in the long run. That is greedy approaches might not be optimal. Reinforcement Learning is an approach where an agent learns how to behave in a environment by performing actions and seeing the results. Reinforcement learning is connected to applications for which the algorithm must make decisions and the decisions bear consequences. The goal is defined by maximization of expected cumulative reward.

The algorithm presents a state, depending on the input data in which a user rewards or punishes the algorithm for the action the algorithm took. The algorithm learns from the reward/punishment and updates itself, this continues [29].

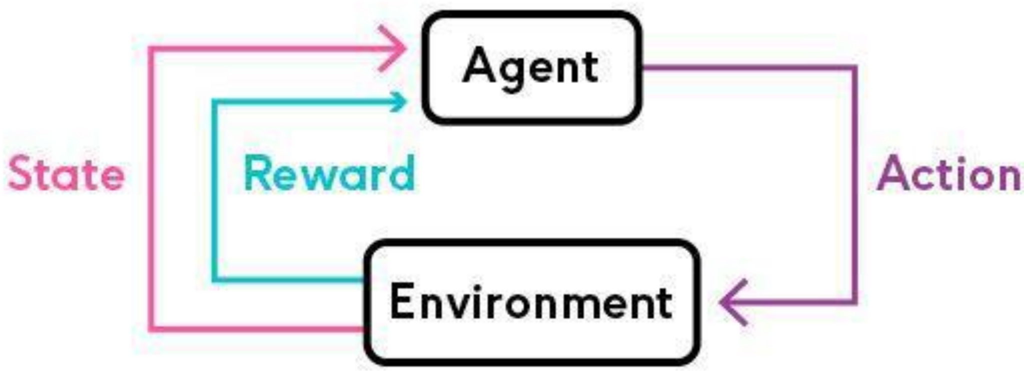

Figure 3.8: Block Diagram of Reinforcement Learning

3.2: Evaluation Parameters:

To understand classifier model's performance, we need to be familiar with some evaluation parameters. A confusion matrix is a table that is used to describe the performance of a classifier algorithm by evaluating the accuracy of it. The elements of confusion matrix are:

True Positive (TP): Which results when classifier model correctly predicts the positive class.

True Negative (TN): Which results when classifier model correctly predicts the negative class.

False Positive (FP): Which results when classifier model incorrectly predicts the positive class.

False Negative (FN): Which results when classifier model incorrectly predicts the negative class.

Predict Positive (0)	True Positive (0)	False Positive (0)
Predict Negative (1)	False Negative (1)	True Negative (1)
	Actually Positive (0)	Actually Negative (1)

Table 3.1: Confusion Matrix

Precision: Precision is a metric that measure how accurate the classifier is, and how many of data returned as correct. A greater precision suggests that there are fewer false positives, whereas a lower precision means that there are more false positives. The ratio of properly categorized examples to total instances is known as precision (P).

$$\text{Precision} = \frac{Truth\ Positive}{Truth\ Positive + False\ Positive} \ldots\ldots\ldots 5$$

Recall: The sensitivity of a classifier is determined by recall, or how much positive data it returns. There are fewer false negatives when the recall is higher. Recall is defined as the proportion of successfully categorized examples to the total number of predicted occurrences.

$$\text{Recall} = \frac{Truth\ Positive}{Truth\ Positive + False\ Negative} \ldots\ldots\ldots\ldots 6$$

F-measure: When precision and recall are combined, the outcome is the weighted harmonic mean of precision and recall, which is known as F-measure.

F-measure = $2*\frac{Precision*Recall}{Precision+Recall}$…………….7

Accuracy: This is the percentage of times the classifier was right. We may achieve this by multiplying the total observations by the sum of the TP and the TN.

Accuracy = $\frac{Truth\ Positive+Truth\ Negative}{Truth\ positive+Truth\ Negative+False\ Negative+False\ Positive}$………8

Chapter: 4(Introduction to Fake Review Detection)

4.1 Fake Review Detection

Online platforms have radically transformed the way people express themselves in the modern internet era. One approach to do so is to express one's thoughts in a product or service review. When making vital decisions, people rely on the analyses that are now available from multiple web sources. They gain a better understanding of the options before selecting. Knowing the difference between fake and genuine reviews is critical since it affects businesses and potential customers. For example, The amount of people using email is rising quickly, and spam email puts a great deal of people's privacy at risk. Spam email can be malicious as well as used for commercial purposes, such as marketing. Numerous machine learning (ML) algorithms exist to identify and analyze spam and unwanted emails [30].

4.1.1 Review Understandings

Online reviews are simply comments, tweets, postings, and opinions shared on various online platforms such as news websites, e-commerce websites, review websites, and social networking platforms. Reviews are defined as an individual's opinion or experience with a product or service [31]. The number of customer reviews has significantly increased in recent years, which has an impact on the potential consumer. Customers' views and experiences with a product or service are also having an increasing influence on marketplaces. Customers' feedback is increasingly being used to transform businesses by improving their goods, services, and marketing offerings [32].

4.1.2 Recognizing Fake Reviews

Customers are free to express their opinions about any business or product on numerous social media platforms such as Facebook, Twitter, and practically all e-commerce websites. Some businesses unfairly advertise their products or services and occasionally insult other businesses due to a lack of laws. Many businesses may hire someone to post positive reviews of their own products to sell those things, while also requesting that individual to write negative reviews of their competitors' products.

Reviews provided by people who have not used the products are considered false [32]. As a result, anyone who posts fraudulent reviews is labeled a spammer [32]. When spammers work together to achieve a shared goal, they are referred to as a spam group [32]. False reviews are an evil behavior known as opinion spamming. Opinion spamming attempts to mislead review readers [1].

4.1.3 The role of sentiment analysis in the detection of fake reviews

Sentiment analysis is a great tool for detecting fraudulent reviews and customer sentiment (SA). Sentiment Analysis (SA), a Natural Language Processing (NLP) technique, is used to collect user sentiments and opinions about any produced objects or supplied services [33]. Sentiment analysis is another term for opinion mining (OM). This is useful while making judgments. Sentiment Analysis (SA) is a popular technique for detecting false positive and negative assessments in consumer opinions and extracting emotions from them. As part of the sentiment analysis classification process, data is categorized into various categories. These categories can be categorical in nature, such as positive or negative, or they can include multiple subcategories, such as happy, sad, angry, and so on [34].

4.1.4 The significance of detecting fake reviews

Because of the rapid growth of the e-commerce industry, user evaluations have become an important source of information for all businesses. User reviews are now extremely significant for a company's capacity to make money. Because many people nowadays strongly rely on reading reviews posted by other users before proceeding with this.

This is why other new users rely on these reviews for knowledge. These reviews have a direct impact on the company's profitability and reputation. False reviews affect the utility, accuracy, and effectiveness of online product reviews [35]. These also undermine the trustworthiness of reviews. To maximize their financial profits, many internet merchants publish positive evaluations for their own business while also posting negative reviews of their competitors' enterprises. These types of erroneous information have a negative influence on profits and mislead customers. As a result, detecting and recognizing fraudulent reviews.

4.2 Antecedent and Consequent of Fake Reviews

Opportunity seeking could be a valid explanation for why spammers create false reviews in that situation. Misleading information includes information about products, such as hoaxes on collaborative platforms, fraudulent reviews in e-commerce, and misleading news on social media (Pantano, 2020) [35]. The issue is rapidly increasing in research, as evidenced by an increase in the number of studies linked to fraudulent reviews that have been published. However, the majority of this research focuses on detection, whereas little is known about the causes and effects [35].

4.2.1 Antecedent of Fake Review

According to studies [35] online product reviews influence customer purchasing decisions (Heydari et al., 2015), as do product reputations, sales volumes, and merchant profitability (Petrescu et al., 2018). Most fake reviews are posted by small business owners, management firms, poor brands, low ratings, and poor quality. But only on rare occasions. In the face of fierce rivalry, dominant brands with high ratings, greater quality, and competitive advantages may give false reviews [35]. Individual customers may submit false reviews to receive incentives, according to Thakur et al. (2018). This behavior is generally motivated by psychological needs that stem from three distinct sources: unhappy consumers, self-appointed brand managers, and social status [35]. There is yet to be a comprehensive study that examines the elements that determine why and how people, review sites, and AI bots create fraudulent reviews. It is also vital to investigate the underlying psychological factors that enable people to write fake reviews without getting outside financial incentives [35]. Furthermore, platforms that publish fraudulent reviews have uncertain intentions and explanations, and AI agents play a role in posting fake reviews [35].

4.2.2 Consequent of Fake Review

Many theoretical models are used to show how fake reviews hurt people, which has caused a lot of worry [35]. Because of fake reviews, the number of online product reviews is always going up (Petrescu et al., 2018). (Luca & Zervas, 2016) [35] Few fake reviews are fair; most are either positive or negative. The wide range of review scores is made worse by fake reviews, which are not like other online product reviews. Customers are tricked into buying things because of fake reviews, which should hurt sales and revenue [35]. But more research needs to be done, and the effects of fake reviews on how online reviews have changed over time should be measured [35].

4.3 Source of Fake Reviews

There are several different kinds of faulty or fake reviews. Here are a few typical locations where most fake reviews are found [36]:

- International vendors: Some service providers or vendors offer both positive and negative ratings to businesses all around the world. Essentially, finding these reviews involves searching for or purchasing reviews from different online platforms.
- Marketers and entrepreneurs: Marketers either directly or indirectly contribute to favorable or unfavorable evaluations. Most of the time, they produce favorable reviews for themselves in order to benefit from them, while producing negative reviews for competitors in order to harm those organizations' businesses.
- Employees: Current employees frequently post positive ratings for their employers in order to receive further advantages, whereas former employees frequently publish negative evaluations of their current employers in revenge for being fired or laid off.
- Friends and family: They occasionally write positive reviews on behalf of a company or brand they are personally linked with, and they periodically publish negative evaluations about the rival of their former employer.
- Customers: On occasion, customers post negative reviews in an effort to get a discount or another type of payment. They genuinely fabricate negative experiences or tell lies about them in order to get discounts or other benefits.

4.4 Methods of Fake Review Detection

Since 2007 [37], the problem of identifying fake reviews has been solved. In order for customers to make better choices regarding which merchants to trust when making purchases, the detection of bogus reviews has become a key issue [38]. The following methods are used to spot fake reviews:

4.4.1 Feature Engineering Extraction

One of the most difficult challenges in detecting fraudulent reviews is feature extraction. This is divided across several nodes, including the reviewer who posted the review, the actual review, the product matching function, and the capacity to collect some network-related data.

4.4.1.1 Review Centric Feature

This approach detects fake reviews based on the content of the reviews supplied by reviewers. This method considers several factors, such as review content similarity, capitalization, use of all capital letters, numerals, brand name, similarity between products and reviews, recurrent use of good and bad words in reviews [39], percentages of pronouns, nouns, adjectives, and verbs, lexical validity, lexical diversity, content diversity, active / passive voice, pictures, and links, among others [39].

4.4.1.2 Reviewer Centric Feature

This method is depending on the actions of reviewers [31]. This technique considers all reviews published by users as well as user information [40]. When employing this strategy, the following characteristics are considered: Account age, profile image, number of reviews posted by a single reviewer, maximum rating per day [39], number of shared/helpful reviews, percentage of positive and negative reviews, ratio of different purchases, rating deviation, review length, and other factors [40].

4.4.1.3 Product-Centric Function

This method is largely concerned with product information. This method takes into account features such as a product's price, sales rank, and other criteria [31].

4.4.1.4 Network-Centric Function

This approach is used to collect IP addresses and GPS data. We can see the timestamp and its pattern regardless of the hour, week, or month. As a result, the traffic patterns of sender IP neighborhood density are also visible. They are often close to the network when it is a spam network, and all of these evaluations over a specific time period originate from the same IP neighborhood. We can also tell which device is being used to post the reviews [40].

4.4.2 Sentiment Analysis Methodologies

To practice sentiment analysis, several different topics have been used. Studies on sentiment analysis for news and blogs, product reviews, and movie reviews, for example [41]. This section discusses some of the sentiment analysis approaches available for detecting fraudulent reviews.

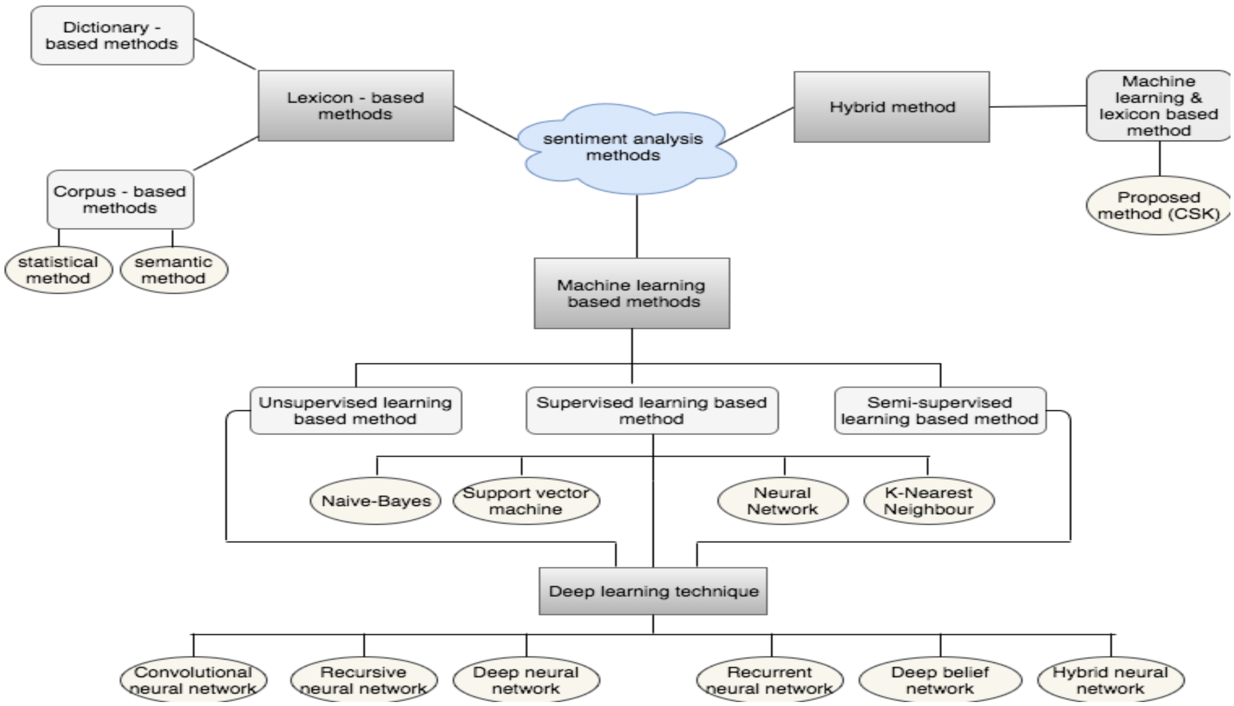

Fig 4.1: Sentiment Analysis Methodologies

[Source Link : https://www.researchgate.net/profile/Seema-Choudhary/publication/326200798/figure/fig1/AS:644988857774082@1530788735622/Techniques-Of-Sentiment-Analysis-31-Machine-Learning-based-TechniqueMachine-Learning.png]

4.4.2.1 Approaches Based on Lexicons

The lexicon-based method determines polarity by comparing sentiment dictionary terms with available data. Manual building, corpus-based methods, or dictionary-based methods can all be used to develop a sentiment lexicon [42]. Manual construction is a difficult and time-consuming task. Opinion words can be generated with a reasonable level of accuracy using corpus-based algorithms. Finally, the idea behind dictionary-based techniques is to first collect a small group of opinion words manually with known orientations, and then to expand this set by searching the WordNet lexicon for their synonyms and antonyms [42].

This technique is managed by a dictionary composed of pre-tagged lexicons. The Tokenizer converts the text provided into tokens [41]. The lexicon of the dictionary is then utilized to match every new token discovered. If there is a successful match, the score is added to the overall score pool of the input text. If the dictionary matches the word "dramatic" positively, the overall score of the text rises. Otherwise, the term is marked as negative, and its score is lowered. Even though this technique appears to be unprofessional in nature, its modifications have proven to be beneficial. Lexical analysis has a disadvantage: as the size of the dictionary (number of words) grows exponentially, so does its performance (in terms of time complexity and accuracy) [41]. Figure depicts the operation of a lexical method.

4.4.2.2 Machine learning-based approaches

Machine learning is one of the most well-liked approaches, drawing interest from researchers due to its scalability and precision. Sentiment analysis makes use of the majority of the supervised learning variations of this method. It involves these three stages: The procedure includes the following steps: data collection, preprocessing, training data, classification, and result plotting [41]. The training data contains a collection of tagged corpora. The classifier receives a set of feature vectors derived from the prior data. A model is created based on the training data set to categorize new or previously unread material. In order for a classifier to be accurate when employing a machine learning technique, the appropriate attributes must be picked [41]. Unigrams (single word phrases), bi-grams (two consecutive phrases), and trigrams are frequently seen in feature vectors (three consecutive phrases). A few of the suggested features include the quantity of positive and negative words, the size of the document, Support Vector Machines (SVM), Naive Bayes (NB), neural networks, and the K-Nearest Neighbour (K-NN) algorithm. Accuracy can range from 63% to 80% depending on the combination of different features used [48]. The following figure (Figure) illustrates the steps involved in machine learning approaches:

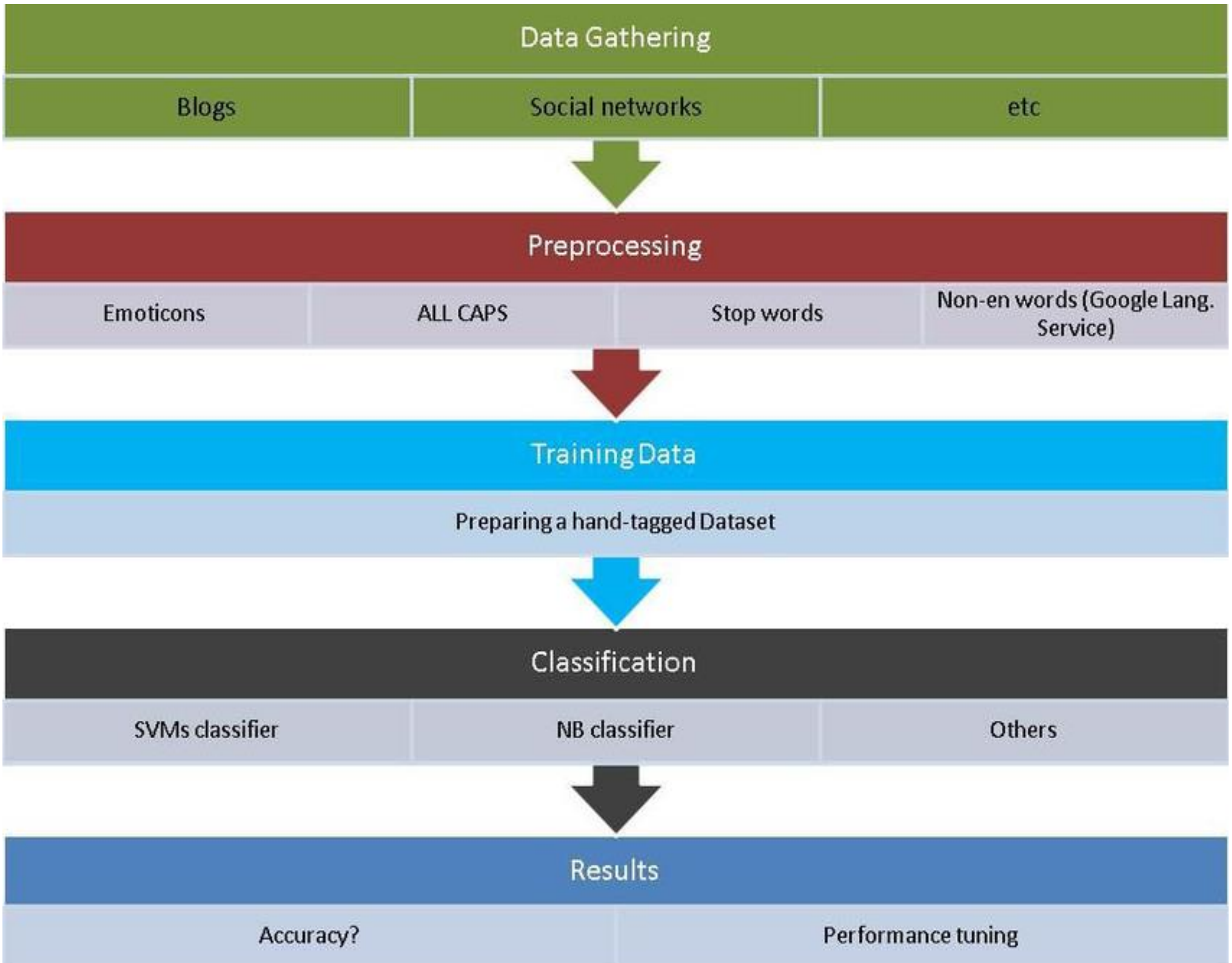

Figure 4.2: Machine learning-based approaches

[Source Link: https://www.researchgate.net/profile/Harsh-Thakkar-11/publication/285648161/figure/fig1/AS:339645796765716@1457989276865/Steps-involved-in-the-machine-learning-approach.png **]**

Regarding classifier design, accessibility of training data, and precise phrase interpretation, machine learning has some limitations. It circumvents the lexical approach's performance degradation restriction and keeps working effectively as the dictionary's size grows exponentially [41].

4.4.2.3 Hybrid Methodologies

By combining both lexicon-based and machine learning methodologies, the performance of sentiment categorization may be improved in the hybrid approach [42]. This methodology may show both the quickness of a lexical approach and the accuracy of a machine learning technique [41]. Adopting these varied approaches has benefits and limitations depending on the purpose of the inquiry. The fundamental advantages of hybrid approaches are the lexicon/learning symbiosis, the detection and assessment of sentiment at the concept level, and the decreased vulnerability to changes in topic domain. Due to the method's major limitation—its inability to discern sentiment—reviews containing a lot of noise (words unrelated to the review's topic) usually receive a neutral score [42].

Chapter:5 (Research Methodology)

As one of the biggest social media platforms in existence today, Facebook is seeing a rise in the popularity of restaurant reviews. We used the Bengali Facebook comments dataset from the largest Bangladeshi Facebook food review community, "Foodbank," to evaluate if comments are good or negative. This section of our thesis will detail how the works were implemented using different strategies. After acquiring a Bengali Facebook dataset from the following Kaggle source, we modified it to meet our requirements [43].Here, the dataset has the following organizational structure: "Speech" is referred to as the Facebook comments made by users who shared their thoughts about the meals they had at various eateries. "Label" is a public sentiment indicator; if it is zero, then the public has something terrible to say about the meal, and if it is one, then they have something nice to say about it.

	Speech	Label	Tag
0	ব্যাস্ততম জীবনের একটু মনরম পরিবেশে সময় কাটা...	1	Positive
1	অসাধারণ পরিবেশ! খুব সুন্দর মার্জিত এবং আন্তরিক...	1	Positive
2	খাবারের মান যথেষ্ট খারাপ। দামের তুলনায় পরিমান ...	0	Negative
3	ভালো বহুত খেয়েছি আমার বাসা থেকে একটু দূরে	1	Positive
4	আমি খাইসি খুবই মজার	1	Positive
5	দয়াল ভান্ডার মিষ্টি চমচম সব গুলা জিনিস আমার অন...	1	Positive
6	মাংস অনেক শক্ত ছিল।আহামরি কিছুই পেলাম না।নরমাল...	0	Negative
7	একদম ভালো না।নিজের অভিজ্ঞতা থেকে বলছি।খুলনার আ...	0	Negative
8	খেয়েই রিগ্রেট করসি, আমি আব্বাস/কামরুল দুইটারই ...	0	Negative
9	কাচ্চি ভাই এর চুই এর থেকেও বাজে টেস্টের এটা। ত...	0	Negative
10	মিরপুরের টায় খেয়েছিলাম, আগুন দামে আমাদের গরিবদ...	0	Neutral

Figure 5.1: Snapshot of a small sample from the Facebook Comment Dataset.

In this thesis, we have attempted to identify the positive and negative Bengali comments on cuisines using a few machine learning techniques. These methods include of multinomial neural networks (MNB), logistic regression (LR), decision trees (DT), random forests (RF), K-nearest neighbors (KNN), and linear support vector machines (SVM).

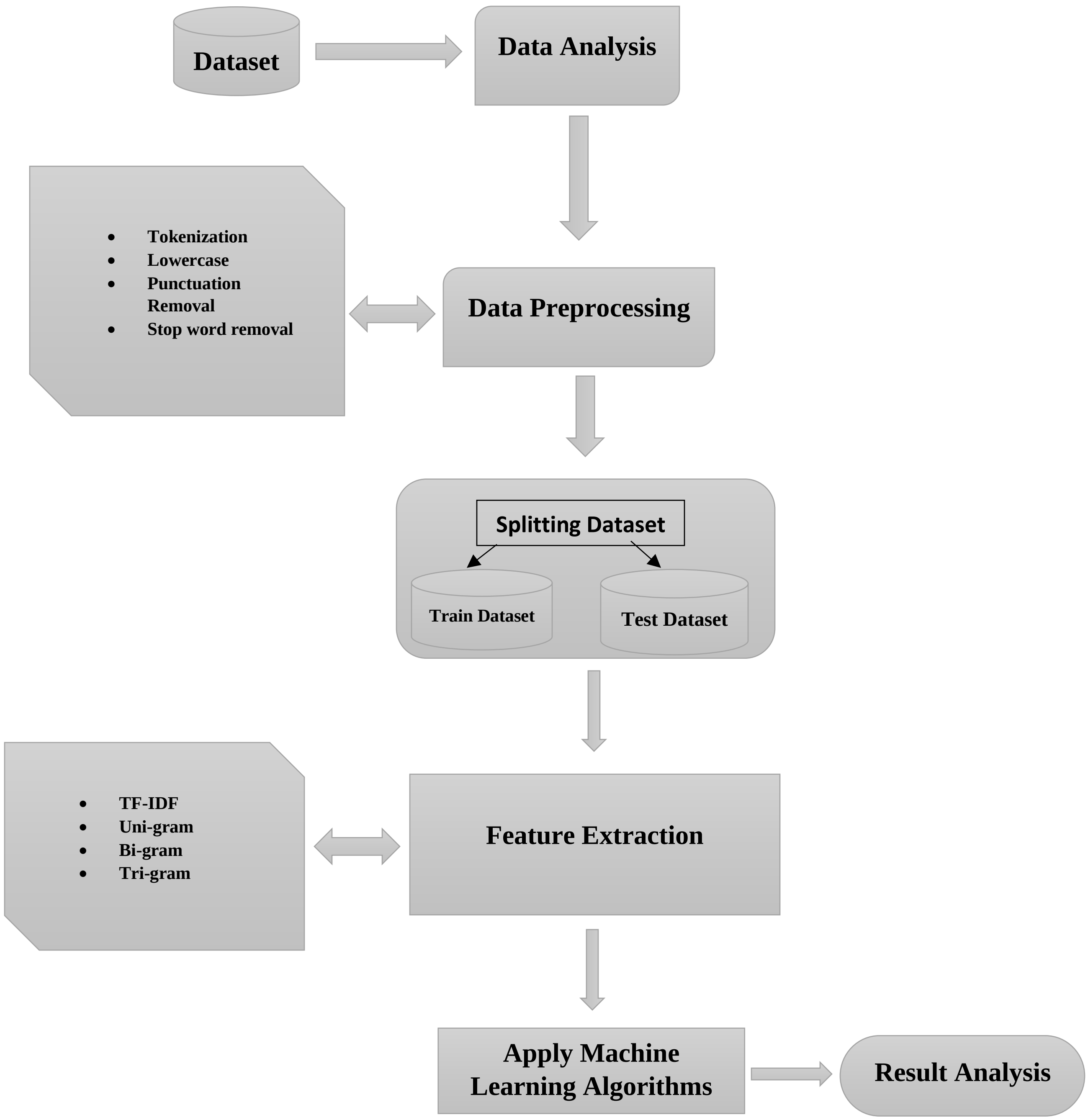

Figure 5.2: Proposed Work model for the Bengali Fake Food Review Detection

5.1: Dataset & Data Analysis

Firstly, we will discuss about the Bengali comments about food which is created by [43]. It includes around 1109 total food reviews from the Foodbank Facebook group, of which there are 563 positive food reviews, 538 negative food reviews, and 8 neutral food reviews. The breakdown of these reviews is displayed in the figure below.

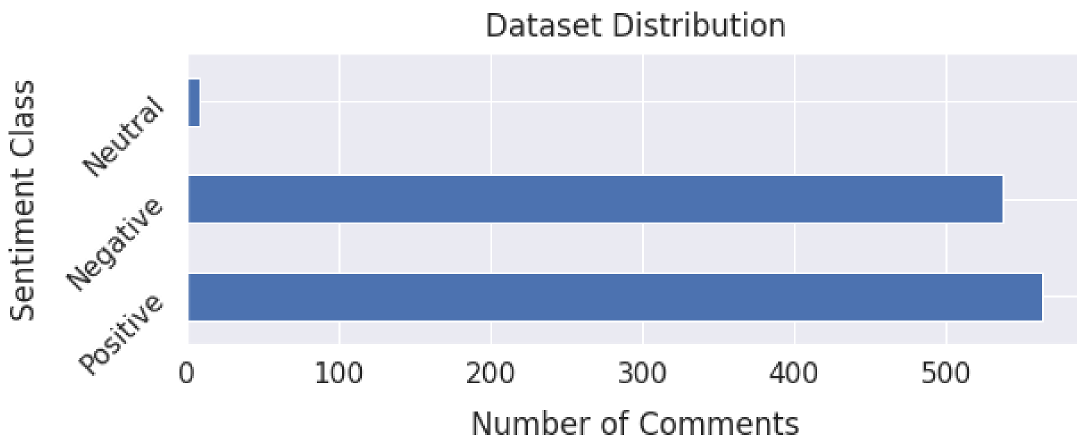

Figure 5.3: Dataset distribution

The percentages of positive reviews, negative reviews, and neutral reviews are as follows: 50.77% for positive reviews, 48.51% for negative reviews, and 0.72% for neutral reviews, which is also shown as a bar graph.

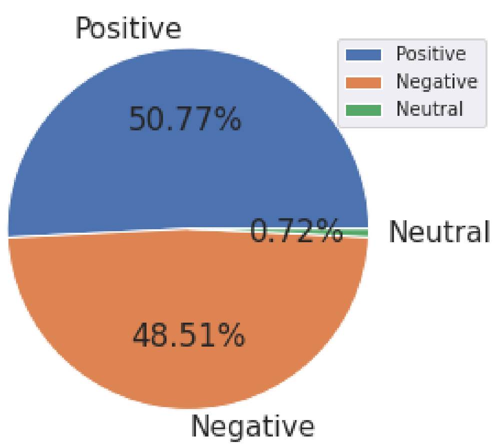

Figure 5.4: Percentage of total reviews

After that, we used data cleaning to take out extraneous punctuation, making the data more appropriate for machine learning techniques.

The dataset's summary was then displayed in three array formats (Documents, Words, and Unique Words), with each format's class distribution within the dataset verified. There are 555 documents covering the topic of the positive, 5552 words, and 1650 unique words. Similarly, there are 534 documents, 7141 words, and 2055 unique words for negative. The total number of documents is eight, while the total word count is ninety-two, and the number of unique words is seventy-eight for neutral. There is a clear proportion between the three categories, as seen in the following diagram.

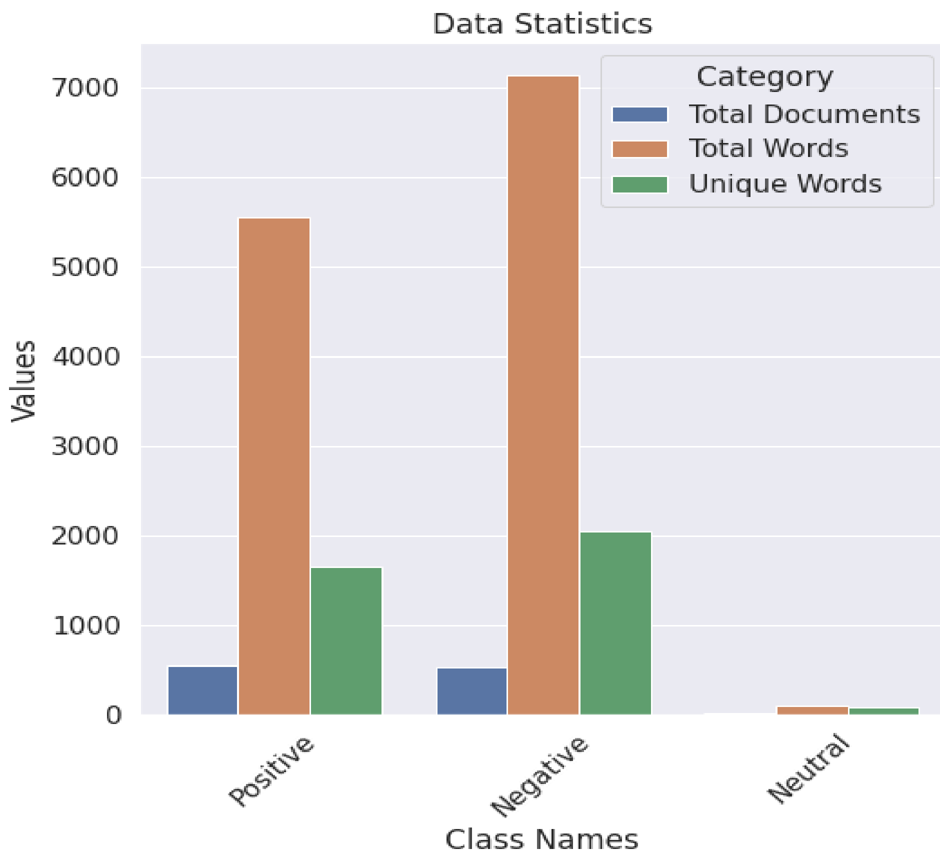

Figure 5.5: Informational Dataset Summary Statistics.

5.1.1: Most Used Words:

In the below figure these are the most frequently occurring words throughout all evaluations. Our sentiment analysis includes all of those terms. It's time to split hairs and separate the positive and negative terms for use in our sentiment study.

Figure 5.5: Word Cloud view for most used words.

These are all the terms that were used throughout the whole dataset to provide either positive or negative reviews of different cuisines. Word cloud was used to illustrate them in the following figure for your convenience.

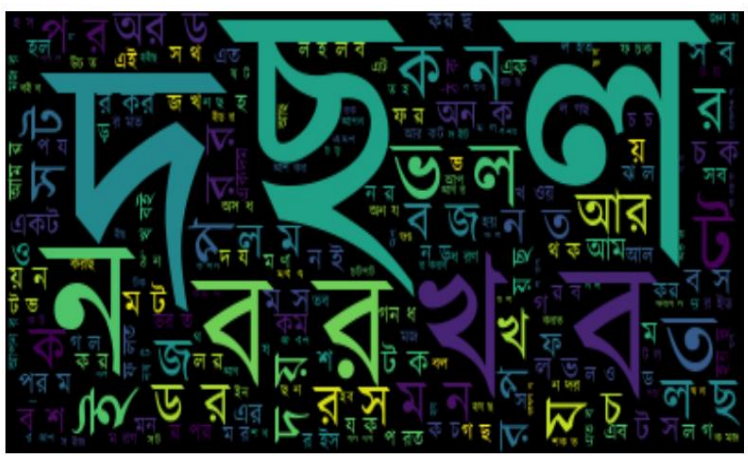

Figure 5.6: Word Cloud view for all words used in reviews.

The illustration that follows presents a word cloud perspective that brings together all positive, negative, and neutral words in one place, displaying them against a varied backdrop in order to facilitate greater comprehension.

Figure 5.7: Word Cloud view in different background

5.2: Data Preprocessing

Given the fact that we are working with a Bangla dataset, the data pre-processing phase will involve us keeping the review text column and the flagged column while removing all of the other columns.

5.2.1: Tokenization:

Tokenization is the process of separating individual words from a string of text. Depending on the context, a token could be a single letter, a few letters, a few digits, or even an entire word.

This is a fundamental challenge in natural language processing (NLP) because every language has its unique grammatical structures, the rules for which can be tricky to codify [44].

The tokenization is depicted in the following figure.

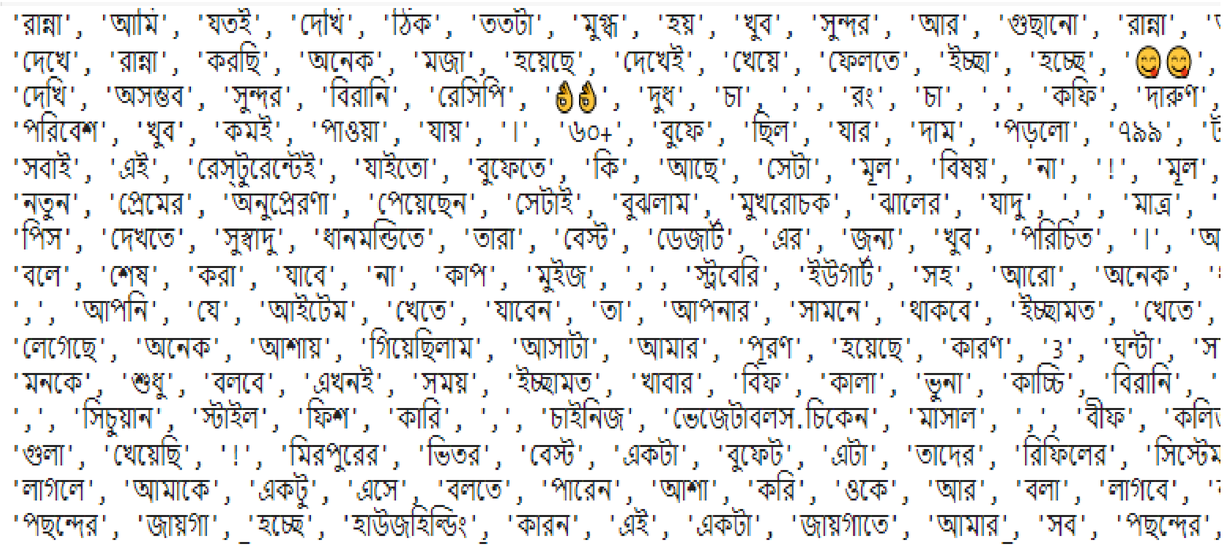
'রান্না', 'আমি', 'যতই', 'দেখি', 'ঠিক', 'ততটা', 'মুগ্ধ', 'হয়', 'খুব', 'সুন্দর', 'আর', 'গুছানো', 'রান্না',
'দেখে', 'রান্না', 'করছি', 'অনেক', 'মজা', 'হয়েছে', 'দেখেই', 'খেয়ে', 'ফেলতে', 'ইচ্ছা', 'হচ্ছে', '😋😋',
'দেখি', 'অসম্ভব', 'সুন্দর', 'বিরানি', 'রেসিপি', '👌👌', 'দুধ', 'চা', ',', 'রং', 'চা', ',', 'কফি', 'দারুণ',
'পরিবেশ', 'খুব', 'কমই', 'পাওয়া', 'যায়', '।', '৬০+', 'বুফে', 'ছিল', 'যার', 'দাম', 'পড়লো', '৭৯৯',
'সবাই', 'এই', 'রেস্টুরেন্টেই', 'যাইতো', 'বুফেতে', 'কি', 'আছে', 'সেটা', 'মূল', 'বিষয়', 'না', '!', 'মূল',
'নতুন', 'প্রেমের', 'অনুপ্রেরণা', 'পেয়েছেন', 'সেটাই', 'বুঝলাম', 'মুখরোচক', 'ঝালের', 'যাদু', ',', 'মাত্র',
'পিস', 'দেখতে', 'সুস্বাদু', 'ধানমন্ডিতে', 'তারা', 'বেস্ট', 'ডেজার্ট', 'এর', 'জন্য', 'খুব', 'পরিচিত', '।',
'বলে', 'শেষ', 'করা', 'যাবে', 'না', 'কাপ', 'মুইজ', ',', 'স্ট্রবেরি', 'ইউগার্ট', 'সহ', 'আরো', 'অনেক',
',', 'আপনি', 'যে', 'আইটেম', 'খেতে', 'যাবেন', 'তা', 'আপনার', 'সামনে', 'থাকবে', 'ইচ্ছামত', 'খেতে',
'লেগেছে', 'অনেক', 'আশায়', 'গিয়েছিলাম', 'আসাটা', 'আমার', 'পূরণ', 'হয়েছে', 'কারণ', '৩', 'ঘন্টা',
'মনকে', 'শুধু', 'বলবে', 'এখনই', 'সময়', 'ইচ্ছামত', 'খাবার', 'বিফ', 'কালা', 'ভুনা', 'কাচ্চি', 'বিরানি',
',', 'সিচুয়ান', 'স্টাইল', 'ফিশ', 'কারি', ',', 'চাইনিজ', 'ভেজেটাবলস.চিকেন', 'মাসাল', ',', 'বীফ',
'গুলা', 'খেয়েছি', '!', 'মিরপুরের', 'ভিতর', 'বেস্ট', 'একটা', 'বুফেট', 'এটা', 'তাদের', 'রিফিলের',
'লাগলে', 'আমাকে', 'একটু', 'এসে', 'বলতে', 'পারেন', 'আশা', 'করি', 'ওকে', 'আর', 'বলা', 'লাগবে',
'পছন্দের', 'জায়গা', 'হচ্ছে', 'হাউজহিল্ডিং', 'কারন', 'এই', 'একটা', 'জায়গাতে', 'আমার', 'সব', 'পছন্দের',

Figure 5.8: Tokenization

5.2.2: Clean Text:

The technique of removing punctuation will aid in treating each text equally. When punctuation is removed, the words data and data! for instance, are regarded similarly. To clean the data, we employ the following techniques:

- Remove all occurrences of the pattern,
- the @ symbol,
- the hashtag symbol,
- the short word symbol,
- the article symbol,
- the white space separator
- the digit separator.
- deleting the username

Because the contraction words will lose their meaning after the punctuation has been removed, we must be careful not to damage the text. Depending on the use cases, we also need to be very attentive when selecting the list of punctuation that we want to omit from the data [45]. Before and after eliminating punctuation, the graph below depicts the real comprehension of the sentence.

```
Original:
 এক কথায় জঘন্য। তিন টুকরো গরুর চর্বি দিয়েছে আর কেমন যেন বাজে গন্ধ
Cleaned:
 এক কথায় জঘন্য  তিন টুকরো গরুর চর্বি দিয়েছে আর কেমন যেন বাজে গন্ধ
 Sentiment:--  Negative

Original:
 হোটেলের চেয়ে ভালো হবে আপনার রেসিপি। অনেক ধন্যবাদ
Cleaned:
 হোটেলের চেয়ে ভালো হবে আপনার রেসিপি  অনেক ধন্যবাদ
 Sentiment:--  Positive

Original:
 মোটা মামার ভেলপুরি+পানিপুরি খিলগাও এর মধ্যে বেস্ট এক কথায় 😅
Cleaned:
 মোটা মামার ভেলপুরি পানিপুরি খিলগাও এর মধ্যে বেস্ট এক কথায়
 Sentiment:--  Positive
```

Figure 5.9: Before and After removing unnecessary punctuations.

5.2.3: Removing Stop words:

One of the most typical preprocessing techniques used in NLP applications is the elimination of stop words. The plan is to simply get rid of words that are used frequently throughout the entire corpus of publications. Articles and pronouns are examples of what are known as "stop words." Some natural language processing tasks, such as information retrieval and classification, have little use for these terms, and hence, they are not very discriminative. On the other hand, stop word removal may have little effect in some NLP uses. The stop word list for any particular language is often a hand-curated list of the most frequently occurring terms in all available corpora. Here, we have removed stop words for Bengali Language which is given below [46].

```
{'নানা', 'একই', 'হইয়া', 'যাওয়ার', 'হওয়া', 'নিজে', 'করি', 'সময়
'মতোই', 'কেউই', 'মতো', 'হওয়ার', 'নেওয়ার', 'দুটো', 'ধরে', 'ক
'তাঁর', 'যারা', 'থাকে', 'থেকেও', 'খুব', 'কোনো', 'পাওয়া', 'সেটা'
'করা', 'চেয়ে', 'ছাড়াও', 'যদি', 'তুমি', 'হয়ে', 'হোক', 'কয়েক',
'তারপর', 'ওদের', 'পর', 'মধ্যে', 'যায়', 'যিনি', 'করেই', 'নতুন',
'তাঁকে', 'কিন্তু', 'যাঁর', 'করবেন', 'কিংবা', 'পেয়ে', 'এতে', 'করতে
'মধ্যেই', 'আপনার', 'পারে', 'দেখা', 'এস', 'ফের', 'পারি', 'অতএ
'সবার', 'ওখানে', 'কথা', 'যাতে', 'সেখানে', 'করে', 'কাজ', 'তাতে
'দিয়েছে', 'নিজেদের', 'উঁনি', 'নিয়ে', 'টি', 'করিয়ে', 'এঁদের', 'তে
'এবং', 'হিসাবে', 'কাছ', 'দেওয়ার', 'হয়েছিল', 'তার', 'গিয়ে', 'ও'
'বিশেষ', 'করলেন', 'তোমাদের', 'দু', 'লক্ষ', 'তেমন', 'হয়তো', 'এ
'হয়েই', 'কোটি', 'করিয়া', 'কেমনে', 'পরেই', 'করেন', 'ইত্যাদি', '
'আমার', 'দেওয়া', 'করার', 'ওর', 'মনে', 'হাজার', 'থাকবেন', 'অ
'যেখানে', 'হয়েই', 'এমনকি', 'ছাড়াও', 'আগেই', 'হইবে', 'হতে', '
'চেয়ে', 'দেওয়া', 'তাকে', 'মোর', 'কত', 'হইয়া', 'সেটাও', 'হবে',
'এখনও', 'কখনও', 'যাওয়া', 'উত্তর', 'চায়', 'বিভিন্ন', 'আগামী', '
```

Figure 5.10: Removing Stop words

5.2.4: Stemming:

Stemming is the process of removing a component of a word or reducing a word to its stem or root. It's possible that this doesn't necessarily indicate we're going to reduce a word to its dictionary root.

```
ব্যাস্ততম  জীবনের  একটু মনরম পারবেশে  সময় কাটানোর জন্য ঘুরে দেখতে পারেন, মন খারা
সাথে বাস্তবের কোন মিল নাই। চোখে দেখা যায় না এমন সাইজের চিকেন বল পেলাম। যাদের টাক
লাগে, অনেক বার খাইছি মাংস অনেক শক্ত ছিল।আহামরি কিছুই পেলাম না।নরমাল ই মনে হইছে
পাতাল। খেয়েই রিগ্রেট করসি, আমি আব্বাস/কামরুল দুইটারই চুই খাইসিলাম কয়েকবার। এটার ১%
আগুন দামে আমাদের গরিবদের খেয়ে পোষাবে না ভাই আহামরি কিছুই না, এটা আমার মতামত।
আমারে কম দিছে কোয়ালিটি ভালো ছিলো তবে , ফ্রি একটা কোক দিলে আরো ভালো হতো কাচ্চি
নিয়েছিলাম. ডিম খিচুড়ি খুবই ভালো লেগেছে খিচুড়ি তো তেলে পুরা ডুবায়ে দিসে..... খুব বাজে
ঝাল কারি অর্ডার করেছিলাম,টেস্ট ঠিক ছিল,কিন্তু পুরোটাই হাড্ডি।অল্প কিছু মাংস ছিল,তাও শক্ত।
বানানো জন্য এদের নোবেল দেওয়া উচিত। তোরা ব্যাবসা কেমনে করতেছিস ভাই। সবাই কে বলছি
ধরেই এই রেস্টুরেন্ট থেকে অর্ডার করি, আর কোন দিন অর্ডার করব না। তেহারী তেল এ ভরা।
```

Figure 5.11: Stemming

5.3: Feature Extraction:

Natural language processing, often known as NLP, is a subfield of artificial intelligence that focuses on giving computers the ability to comprehend human languages, either in written or spoken form, by converting them into numerical format vectors. This can be done in either text or voice form. Word Embedding is one of the most common techniques for vectorizing texts since it does not produce a spare matrix of vectorized sentences, thus the computation cost is inexpensive, and it preserves the majority of the linguistic information that is already there in the phrase. In this article, we will go through two of the most common variants of frequency-based Word Embedding:

- Count Vectorizer
- TF-IDF Model

5.3.1: Count Vectorizer:

The word count in a text can be obtained with the help of Count Vectorizer, a minimalist vectorizer. If you have many text documents and want to know how many words and tokens are in them, Count Vectorizer is the tool for you.

Think about the phrase, “The cup is on the table.”

Information = [“The,” “cup,” “is,” “present,” “on,” “the,” and “table”]

A vector representation of this is,

Count Vectorizer						
Words	The	Cup	Is	Present	On	table
Sentence	2	1	1	1	1	1

Table 5.1: Count Vectorizer

Using the Count Vectorizer technique, we can see all the words in the sentence in a single table, with repeated words counted as a single occurrence.

Sklearn's Count Vectorizer is a useful tool for performing count vectorization. Find a working example of its application below. Documents are encoded by the Count Vectorizer according to the frequency with which individual words appear in the text. Word and n-gram frequencies are provided in a sparse matrix format by Count Vectorizer [47].

5.3.2: TF-IDF Model:

A statistical measure for determining which words are most relevant in a text, TF-IDF is an ML method. An individual document or a collection of documents might serve as the source material (corpus). Inverse Document Frequency (IDF) and Term Frequency (TF) are the two components of this statistic (IDF).

The TF-IDF document analysis tool takes into account how often specific words appear in the text. The frequency with which words appear in the papers is measured. Calculating TF involves dividing the frequency with which a given word appears in the document I by the total number of words (N) (j) [48].

TF (i) = log (frequency (i,j)) / log (N (j))…………1

A word's rarity is quantified by its IDF score. Words that appear more frequently in a document receive higher emphasis from TF, therefore this is crucial. Even though they may be underrepresented in the corpus, uncommon words may be a rich source of data. That data is collected by IDF. You can figure it out by dividing N by d, where d is the number of documents and N is the total number of documents (i).

IDF (i) = log (N (d) / frequency (d,i))………….2

To mitigate the weight of extremely high TF and IDF values, the log is used in the aforementioned calculations. By multiplying the TF score by the IDF score, we get the TF-IDF score.

For less complex ML and NLP issues, the TF-IDF algorithm is often the method of choice. Information retrieval, keyword extraction, stop word elimination (including a, the, are, and is), and simple text analysis are its strongest suits. As a result, it is inefficient in capturing the meaning of words in context [48].

If we compare TF-IDF model with Count Vectorizer we can see that, Count Vectorizer shows the total frequency with which a word is used, whereas TF-IDF illustrates the significance of the term in the context of the document. For the feature extraction we have applied TF-IDF in our research work.

5.4: Splitting Training and Testing Data:

We split our dataset into a training set and a testing set for the sake of evaluation. Our machine learning model is trained using 80% of the comprises data and then tested using 20%.

To break it down, there are a total of 887 training reviews and 221 test reviews in the dataset.

Chapter: 6 (Result and Analysis)

We have selected the dataset for our research that include reviews that have been classified as fake and real. In this essay, we investigate the relationship between reviews that were actually written and those that were tagged as fake or not. We have 1109 food reviews in the dataset, with 50.77% of them classified as positive and 48.51% as negative and 0.72% as neutral.

Both machine learning and deep learning methods have been used in this investigation.

To identify negative reviews using supervised learning, machine learning classifiers such as Multinomial Naive Bayes, Bernoulli Naive Bayes, Random Forrest, k-Nearest Neighbors, Logistic Regression, Support Vector Machine (Linear), and Decision Tree are used. We used TF-IDF for feature extraction because it produces the best results for both of our datasets. We used CNN deep learning techniques to identify fake reviews. Below are the results of our experiment using machine learning classifiers:

Table 6.1: Bangla Food Review Dataset

ML Classifier	Feature Matrix	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression	Uni-gram	90.37%	0.83	0.95	0.88
	Bi-gram	86.24%	0.76	0.97	0.85
	Tri-gram	82.57%	0.70	0.98	0.82
Decision Tree	Uni-gram	83.03%	0.73	0.91	0.81
	Bi-gram	86.24%	0.76	0.95	0.87
	Tri-gram	82.57%	0.72	0.90	0.80
Random Forest	Uni-gram	86.24%	0.76	0.97	0.85
	Bi-gram	84.86%	0.74	0.97	0.84
	Tri-gram	84.40%	0.73	0.98	0.83

Multinomial Naïve Bayes	Uni-gram	88.53%	0.82	0.91	0.86
	Bi-gram	87.61%	0.80	0.92	0.86
	Tri-gram	86.24%	0.78	0.92	0.84
K-Nearest Neighbor	Uni-gram	82.11%	0.74	0.86	0.79
	Bi-gram	83.49%	0.76	0.86	0.81
	Tri-gram	83.94%	0.77	0.85	0.81
Linear Support Vector Machine	Uni-gram	83.79%	0.71	0.98	0.83
	Bi-gram	62.39%	0.51	1.00	0.68
	Tri-gram	43.58%	0.41	1.00	0.59
RBF Support Vector Machine	Uni-gram	87.16%	0.78	0.95	0.85
	Bi-gram	81.19%	0.68	1.00	0.81
	Tri-gram	63.76%	0.52	1.00	0.69

Tokenization, lowercase, stop words removal, and stemming were all used in this data pre-processing step. We used the term frequency-inverse document frequency (TF-IDF) method to extract features. By implementing several machine learning techniques,

Includes a unigram, a bigram, and a trigram. A unigram is a string of words taken from a text where each word can appear wherever in the string without reference to the previous work. Each word in a bigram is dependent on the one that comes before it in the phrase. In Trigram, each word's appearance is tied to those of two others.

Applying the data preparation model, to begin, our Logistic Regression classifier, which makes use of a single-word feature, has a 90.37% success rate. The total F1-Score, Precision, and Recall are all 0.83, 0.95, and 0.88. Second, we have utilized a bigram, which has an F1-Score of 0.85, a precision of 0.97, and a recall of 0.85, for an overall accuracy of 86.24%. Finally, Trigram was used, and its results showed an accuracy of 82.57%, which was lower than those of both unigram and bigram. Here, we have precision = 0.70, recall = 0.98, and F1-Score = 0.82. Out of the three features used in this approach, the Unigram feature provides the highest level of precision.

We were able to improve the accuracy of the Decision Tree classifier by 83.03% by utilizing the unigram feature. In which the F1-Score, Recall, and Precision all sum to 0.73. Second, we have employed a bigram, which has an F1-Score of 0.87, a precision of 0.76, and a recall of 0.95. Finally, Trigram was used, and its results showed an accuracy of 82.57%, which was lower than those of both unigram and bigram. Here, we have precision = 0.72, recall = 0.90, and F1-Score = 0.80. Bigram feature provides the highest accuracy in this algorithm.

By utilizing the Unigram feature of the Random Forest classifier, we were able to improve its accuracy to 86.24%. In terms of the unigram, the precision, recall, and F1-score each come in at 0.76, 0.97, and 0.85 respectively. Once more, we have made use of the bigram feature, which has an accuracy rate of 84.86% and a precision score of 0.74, a recall score of 0.97, and an F1 score of 0.84. The final step that we took was to use the Trigram feature, which showed an accuracy of 84.40%, which was lower than both the unigram and the bigram. In this case, the scores for precision, recall, and F1-Score are 0.73, 0.98, and 0.83 respectively. In this algorithm, the unigram feature provides the highest accuracy compared to the other two features, despite the fact that their accuracy rates are quite comparable to one another.

We were able to get an accuracy of 88.53% for the unigram feature, 87.61% for the bigram feature, and 86.24% for the trigram feature with the multinomial naive bayes classifier. Whereas Precision, Recall, and F1-Score are three of the qualities that are quite comparable to one another The maximum F1-Score possible is 0.86, the best precision possible is 0.82, and the highest recall possible for both bigrams and trigrams is 0.92.

In the K-Nearest Neighbors classifier we have found the accuracy 82.11% by making use of its Unigram feature. The F1-score, precision, and recall for the unigram are all 0.74. We've used the bigram function again, and it's helped us get an impressive 83.49 percent correct, 0.76 precision, 0.86 recall, and 0.81 F1 score. Finally, we used the Trigram function, which performed worse than both the unigram and the bigram methods combined (83.94% accuracy). Here, the F1-Score, recall, and precision all average out to 0.77. These three features are used in this method and together they give sufficient accuracy.

We found that a Support Vector Machine (linear) classifier with a unigram feature yielded the highest test accuracy (83.49%). The 0.71 precision, 0.98 recall, and 0.83 F1-Score are all record highs for us. The 62.39% test accuracy seen before and after bigram feature is clearly lower. The values for precision (0.51), recall (1.0), and f-1 (0.68) on the test are as follows. When using the trigram function, accuracy drops dramatically to just 43.58%. In addition, its accuracy is the worst of the 0.41 alternatives. With a recall of 1.00 and an F1-Score of around 0.59, this model has poor predictive power.

We found that a Support Vector Machine (RBF) classifier with a unigram feature yielded the greatest test accuracy of 87.16%. Our best-ever values for precision (0.78), recall (0.95), and F1-Score (0.86), respectively, are achieved. The test accuracy decreases to 81.19 percent after implementing bigrams. There is a 0.68 f-1 score, a 100% recall, and a perfect accuracy on the test. Accuracy drops dramatically to just 63.76% while using the trigram option. And its accuracy is likewise the worst of the 0.52s. The F1-Score is roughly 0.69, and the recall is the greatest at 1.00. accuracy is likewise the worst of the 0.52s. The F1-Score is roughly 0.69, and the recall is the greatest at 1.00.

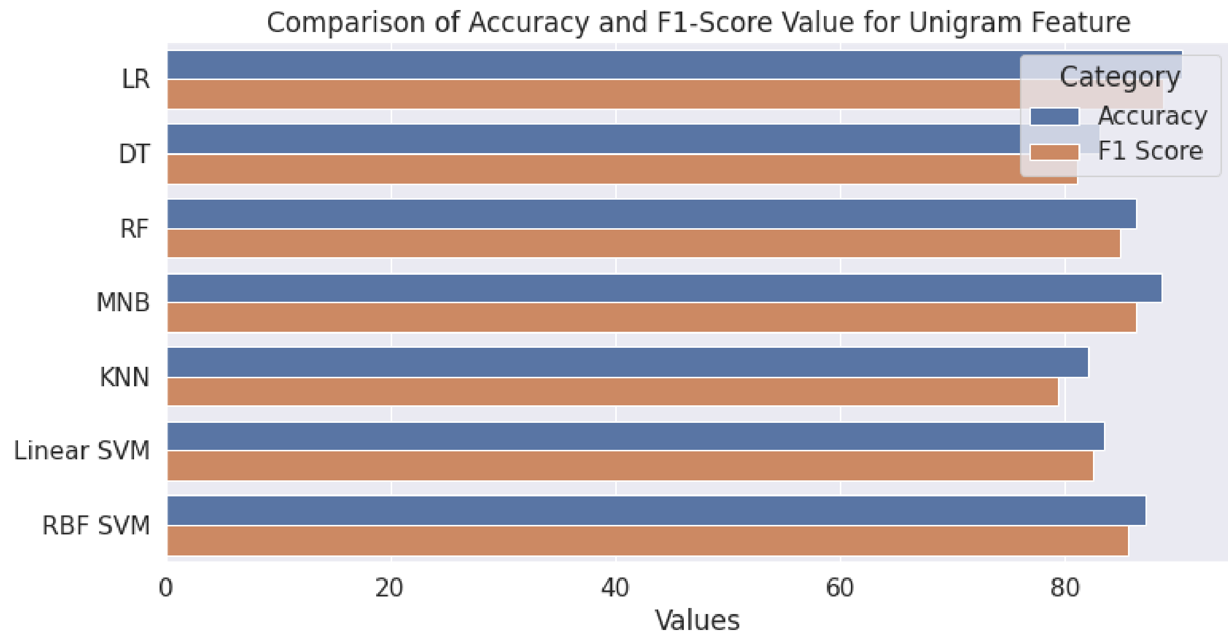

Figure 6.1: The experimental results and comparison for Unigram

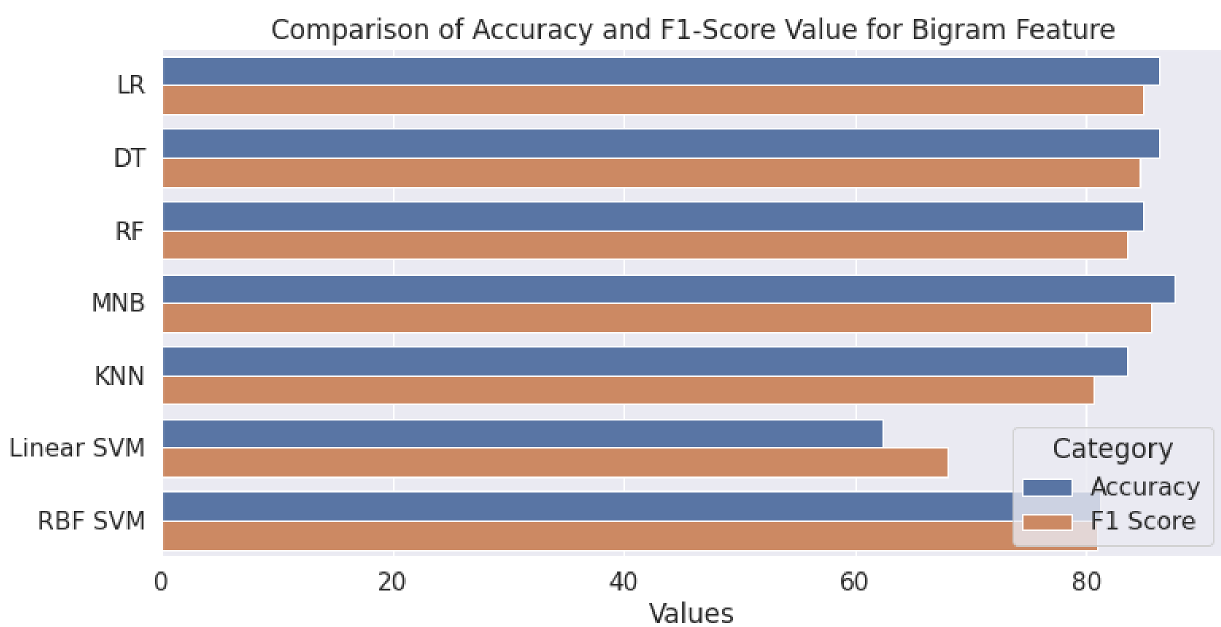

Figure 6.2: The experimental results and comparison for Bigram

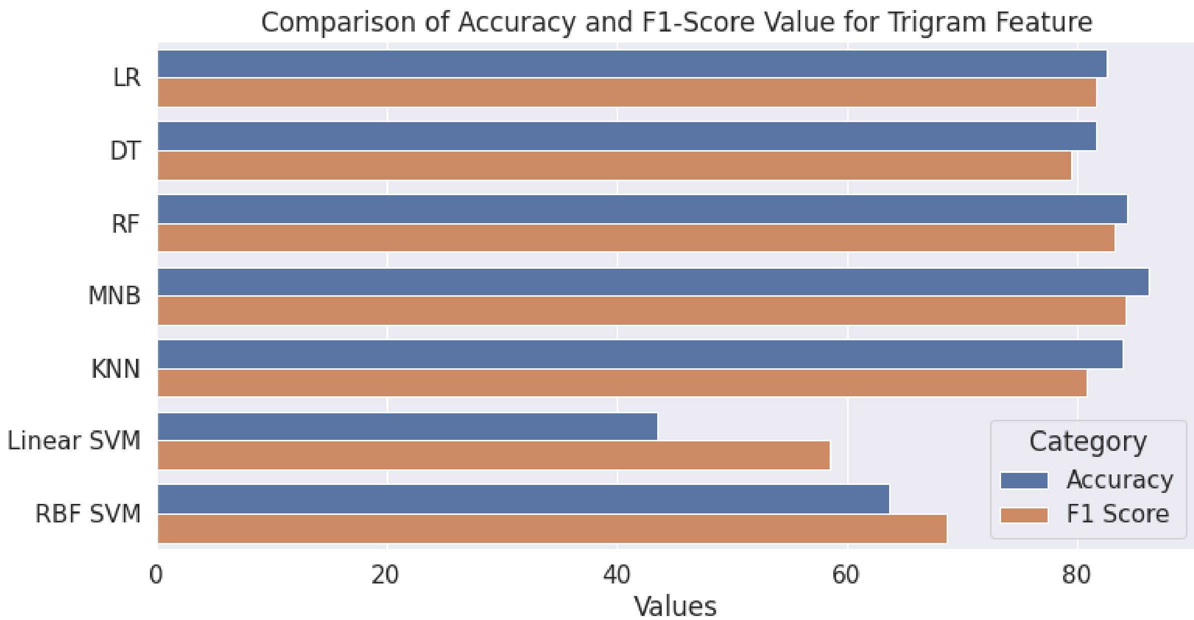

Figure 6.3: The experimental results and comparison for Trigram

Above Figures show that of the seven classifiers tested, K-Nearest Neighbors (KNN) has the lowest accuracy (at 82.11%) and Linear Regression (LR) has the highest (at 90.37%) when utilizing the Unigram feature. Multinomial Naive Bayes (MNB) is the most accurate classifier we have for Bigram, with an accuracy of 87.61%; Linear Support Vector Machine (Linear SVM) is the least accurate, at 62.39%. Finally, Multinomial Naive Bayes (MNB) achieves the maximum accuracy (86.24%) when applied to the trigram feature, whereas Linear Support Vector Machine achieves the lowest accuracy (43.58%) of the tested methods. The highest F1-Score (0.89), highest Precision Score (0.83), and highest Recall Score (0.98 all went to Linear SVM for Unigram classification. As a Bigram, MNB's highest F1-score was 0.86.

At a precision score of 0.80, MNB excelled. The best Recall Score was achieved with Linear SVM at 1.00. MNB's highest F1-score for the Trigram feature was = 0.84. At 0.78, MNB obtained its highest-ever Precision Score. Linear SVM had the highest recall at = 1.00.

6.1 Applying Confusion Matrix for Machine Algorithms

Adding to the Matrix of confusion in our experiment allows for more precise data collection. By using the confusion matrix function, we can see the confusion matrix for each model we tested.

We use the confusion matrix to get a clearer picture of whether or not all of our models are able to determine how many values were predicted correctly or incorrectly. With the help of the Confusion Matrix, we can easily determine how many of the missing values were accurately detected by our models. Furthermore, it has demonstrated that can determine the specific type of error being produced by our model. We have utilized the confusion matrix with many classifier models, including the Support Vector Machine, the Naive Bayes Classifier, the Random Forest, and the Logistic Regression.

The confusion matrix is shown in the below figures,

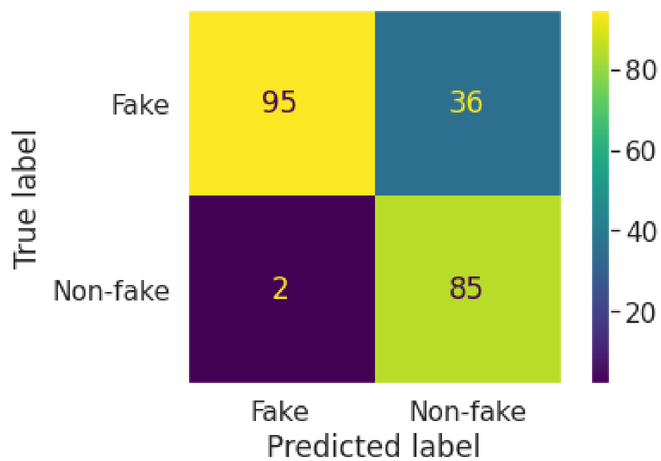

Figure 6.4: Confusion Matrix of LR

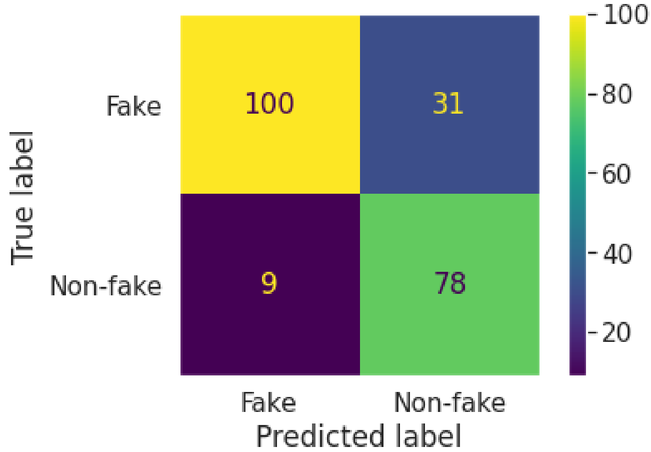

Figure 6.5: Confusion Matrix of DT

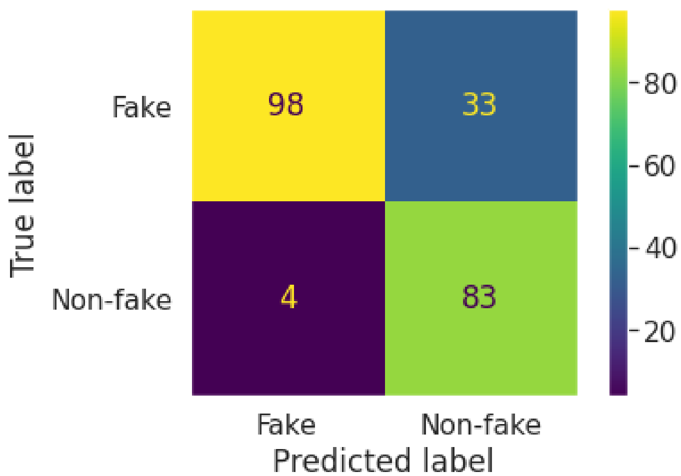

Figure 6.6: Confusion Matrix of RF

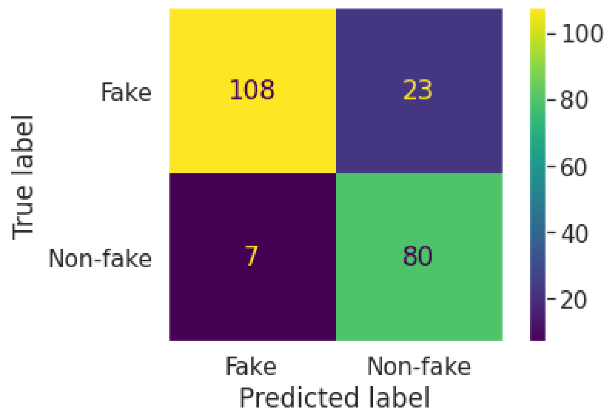

Figure 6.7: Confusion Matrix of MNB

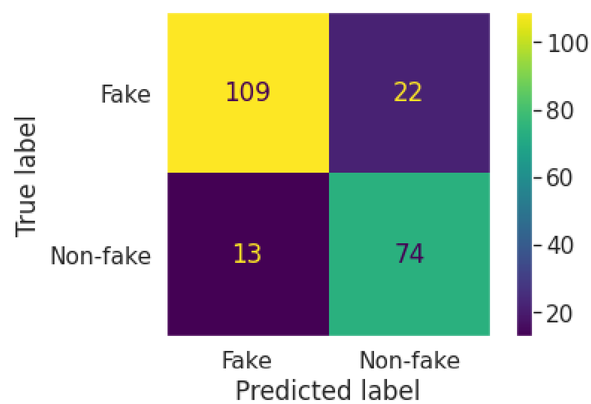

Figure 6.8: Confusion Matrix of KNN

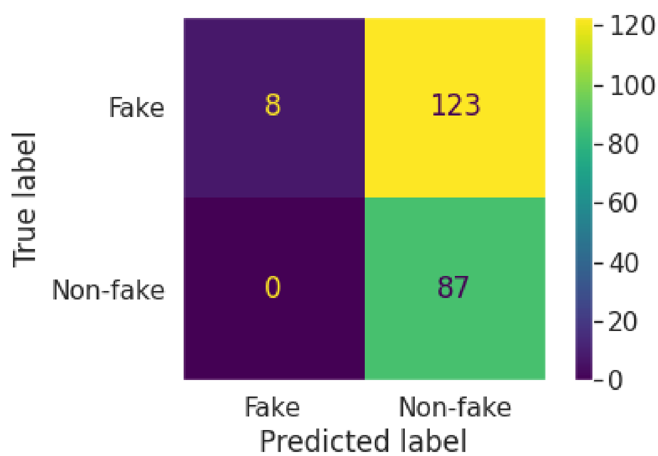

Figure 6.9: Confusion Matrix of LSVM

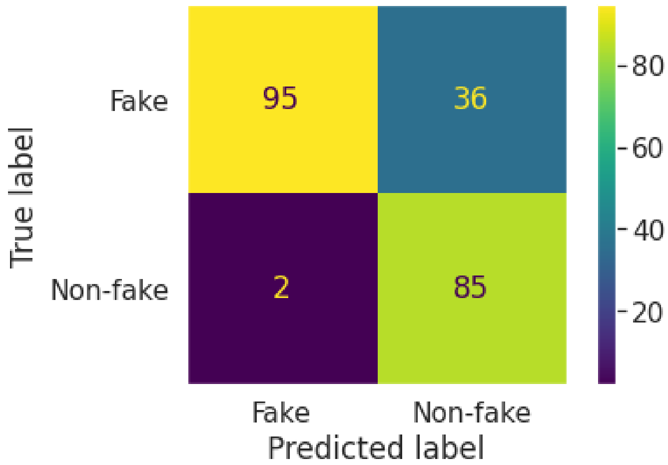

Figure 6.10: Confusion Matrix of RBF SVM

According to the confusion matrices, the classifiers got a large percentage of the fake and non-fake reviews right for the dataset. These results demonstrate that our machine learning models are satisfactory.

Chapter: 7(Conclusion)

7.1 Research Challenges

There are numerous difficulties in identifying fraudulent reviews. From our perspective, the dataset is crucial to achieving higher accuracy. Because in our situation, we obtained the Bangla dataset from Kaggle, updated it, and added numerous data points from the comments relating to food. Working with the Bangla food dataset is really difficult, and we ran into numerous issues when trying to parse the Bangla sentences. We just have a few resources available for our Bangla research projects in NLP. We have improved accuracy overall after applying machine learning and deep learning algorithms to datasets. In addition, there are certain additional difficulties encountered when identifying false reviews, such as the difficulty in doing so when there is only one review available for a given product. Ratings can sometimes make it difficult to tell which are real and which are fraudulent reviews. In certain circumstances, fraudulent reviews are purposefully written exactly in the same manner as genuine evaluations, confusing the reader as to the review's behavior. Additionally, the classification and preprocessing phases had a significant influence on identifying the fraudulent reviews before training and testing even began.

7.2 Future Work

We will continue our research in the future with the goal of enhancing the efficacy of the strategies we have employed. We also intend to suggest some algorithms for identifying bogus reviews. Considering that Bangla-BERT, RNN-LSTM is the most recent natural language processing technique, we are also interested in applying it to our datasets to evaluate how it performs.

Additionally, we wish to expand on this work by conducting a similar study on whole new Bangla datasets. We are one step closer to developing an automated tool for detecting false reviews by categorizing Bangla fraudulent food reviews from social media networks. We also anticipate that this study establishes a baseline for subsequent experiments and broadens the range of available options for the detection of false reviews. The social media data will make sure that the linguistic variances are handled. We would like to delve even deeper, assess the results of this review dissemination, and develop straightforward methods for more accurate prediction.

7.3 Conclusion

In this age of digitalization, when the internet is rapidly evolving, we are forced to interact with a variety of online platforms. These days, most consumers would rather shop online than in a mall, and while making a purchase, they often go to other customers' reviews for guidance. For this reason, most consumers rely heavily on reviews they read online. Because of this, figuring out how to tell a real review from a fraudulent one has become an active and exciting field of study. This thesis work illustrates the use of machine learning techniques like Multinomial Naive Bayes, Bernoulli Naive Bayes, Random Forrest, k-Nearest Neighbors, Logistic Regression, Support Vector Machine (Linear), and Decision Tree to spot fraudulent testimonials. The Logistic Regression yielded the highest unigram output accuracy of 90.37 %. In our dataset, MNB achieves 87.61% accuracy for bigrams and 86.24% accuracy for trigrams. To evaluate how successfully these algorithms construct the model, we have used the Confusion Matrix to them. We have excellent precision, recall, and f-1 score in this example. For the detection of fake food reviews, a machine learning system is the best option.

References

[1] Shivaprasad T K, Jyothi Shetty, "Sentiment Analysis of Product Reviews:A Review," *International Conference on Inventive Communication and Computational Technologies*, pp. 298-303, 2017.

[2] Rami mohawesh, shuxiang xu, son n. tran, Robert ollington, matthew springer, yaser jararweh, sumbal maqsood, "Fake Reviews Detection: A Survey," *Digital Object Identifier*, vol. 9, pp. 65772-65802, 2021.

[3] Elshrif Elmurngi, Abdelouahed Gherbi, "Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques," *The Sixth International Conference on Data Analytics*, pp. 65-72, 2017.

[4] Ahmed M. Elmogy. Usman Tariq, Atef Ibrahim, Ammar Mohammed, "Fake Reviews Detection using Supervised Machine Learning," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, pp. 601-606, 2021.

[5] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, Natalie Glance, "Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews".

[6] Joy Chandra Gope, Tanjim Tabassum, Mir Md. Mabrur, Keping Yu, Mohammad Arifuzzaman, "Sentiment Analysis of Amazon Product Reviews Using Machine Learning and Deep Learning Models," *International Conference on Advancement in Electrical and Electronic Engineering*, 2022.

[7] Nitish Ranjan Bhowmik, Mohammad Arifuzzaman, M. Rubaiyat Hossain Mondal, M. S. Islam, "Bangla Text Sentiment Analysis Using Supervised Machine Learning with Extended Lexicon Dictionary," *Atlantis Press B.V*, vol. 1, pp. 34-45, 2021.

[8] Md. Rakibul Hasan, Maisha Maliha, M. Arifuzzaman, "Sentiment Analysis with NLP on Twitter Data," *International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering*, 2019.

[9] Palak Baid, Apoorva Gupta, Neelam Chaplot, "Sentiment Analysis of Movie Reviews using Machine Learning Techniques," *International Journal of Computer Applications*, vol. 179, no. 7, pp. 45-49, 2017.

[10] Rajkumar S. Jagdale, Vishal S. Shirsat, Sachin N. Deshmukh, "Sentiment Analysis on Product Reviews Using Machine Learning Techniques," *Springer, Singapore,* vol. 768, pp. 639-647, 2018.

[11] R. R. Neethu M S, "Sentiment Analysis in Twitter using Machine Learning Techniques," *ICCCNT,* 2013.

[12] Md. Rafidul Hasan Khan, Umme Sunzida Afroz , Abu Kaisar Mohammad Masum, Sheikh Abujar, Syed Akhter Hossain, "Sentiment Analysis from Bengali Depression Dataset using Machine Learning," *International Conference on Computing and Networking Technology,* 2020.

[13] Rashedul Amin Tuhin, Bechitra Kumar Paul, Faria Nawrine, Mahbuba Akter, Amit Kumar Das, "An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques," *h International Conference on Computer and Communication Systems,* pp. 360-364, 2019.

[14] Minhajul Abedin Shafin, Md. Mehedi Hasan, Md. Rejaul Alam, Mosaddek Ali Mithu, Arafat Ulllah Nur, Md. Omar Faruk, "Product Review Sentiment Analysis by Using NLP and Machine Learning in Bangla Language," *International Conference on Computer and Information Technology,* 2020.

[15] Petr Hajek, Aliaksandr Barushka, Michal Munk, "Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining," *Springer-Verlag London Ltd,* 2020.

[16] Rodrigo Barbado, Oscar Araque, Carlos A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *ELSEVIER,* vol. 56, pp. 1234-1244, 2019.

[17] Mohd. Istiaq Hossain Junaid, Faisal Hossain, Udyan Saha Upal, Anjana Tameem, Abul Kashim, Ahmed Fahmin, "Bangla Food Review Sentimental Analysis using Machine Learning," *Researchgate,* 2022.

[18] Nitish RanjanBhowmik, MohammadArifuzzaman, M. Rubaiyat HossainMondal, "Sentiment analysis on Bangla text using extended lexicon dictionary and deep learning algorithms," *ScienceDirect,* vol. Volume 13, pp. 1-14, 2022.

[19] S. P. Ripa, F. Islam and M. Arifuzzaman, "The Emergence Threat of Phishing Attack and The Detection Techniques Using Machine Learning Models," *IEEE Internet of Things Journal,* 2021.

[20] Masud Rana Basunia, Ismot Ara Pervin , Md. Al Mahmud, Suman Saha, Dr. Mohammad Arifuzzaman, "On Predicting and Analyzing Breast Cancer using Data Mining Approach," *IEEE Conference*, 2020.

[21] D. Johnson, "Supervised Machine Learning: What is, Algorithms with Examples," guru99, [Online].

[22] "Naïve Bayes Classifier Algorithm," Javatpoint, [Online].

[23] "Logistic Regression in Machine Learning," javatpoint, [Online].

[24] A. Saini, "Support Vector Machine(SVM): A Complete guide for beginners," analyticsvidaya, 12 October 2021. [Online].

[25] "Decision Tree," geeksforgeeks.org, 2023 January 10. [Online].

[26] "Random Forest Algorithm," javaTpoint, [Online].

[27] A. ROSS, "WHAT IS UNSUPERVISED LEARNING?," unsupervised, 21 September 2021. [Online].

[28] V. V. PV, "What is Reinforcement Learning?: Components, Approaches, Markov Decision Process, Formalisation using MDP," medium, 2 August 2020. [Online].

[29] K. B. Błażej Osiński, "What is reinforcement learning? The complete guide," deepsense.ai, 5 July 2018. [Online].

[30] T. Toma, S. Hassan and M. Arifuzzaman, "An Analysis of Supervised Machine Learning Algorithms for Spam Email Detection," *IEEE* , 2021.

[31] P. R. P. Nidhi A. Pate, "A Survey on Fake Review Detection using Machine Learning Techniques," *International Conference on Computing Communication and Automation*, pp. 1-6, 2018.

[32] RAMI MOHAWESH, , SHUXIANG XU, , SON N. TRAN, ROBERT OLLINGTON, MATTHEW SPRINGER, YASER JARARWEH, SUMBAL MAQSOOD, "Fake Reviews Detection: A Survey," *SemanticScholar*, pp. 65771-65802, 2021.

[33] K. M. Gurpreet Kaur, "A Comprehensive Overview of Sentiment Analysis and Fake Review Detection," *Springer Singapore*, pp. 293-304, 2021.

[34] "Top Data Science Project – Sentiment Analysis Project in R," Data Flair, 6 February 2022. [Online].

[35] D. S. S. D. S. Uma Maheswari, "Detection of Fake and Genuine Reviews with Hybridization of Fuzzy and Neural Networks Techniques," *Informatica,* pp. 74-95, 2021.

[36] "Tools for Fake Review Detection: How to Spot Fake Reviews," reviewtrackers, 09 November 2022. [Online].

[37] B. L. Nitin jindal, "Opinion spam and analysis," *Proceedings of the 2008 International Conference on Web Search and Data Minin,* pp. 219-230, 2008.

[38] A. Y. C. Snehasish Banerjee, "Applauses in hotel reviews: Genuine or deceptive?," *2014 Science and Information Conference,* pp. 938-942, 2014.

[39] V. Ranganathan, "Fake review detection : framework, challenges and future," Analytics India Magazine, 2020. [Online].

[40] M. M. Ajay Rastogi, "Opinion Spam Detection in Online Reviews," *Journal of Information & Knowledge Management* , pp. 1-38, 2017.

[41] D. P. Harsh Thakkar, "Approaches for Sentiment Analysis on Twitter: A State-of-Art study," *Cornell University,* 2015.

[42] F. F. P. G. T. G. Alessia D'Andrea, "Approaches, Tools and Applications for Sentiment Analysis Implementation," *International Journal of Computer Applications,* pp. 26-33, 2015.

[43] S. PSYCHO, "Sentiment Analysis of Food Reviews on Bengali Text," kaggle, 2022. [Online].

[44] T. Perry, "What is Tokenization in Natural Language Processing (NLP)?," machinelearningplus, 1 February 2021. [Online].

[45] K. Rastogi, "Text Cleaning Methods in NLP," AnalyticsVidhya, 31 January 2022. [Online].

[46] J. P. D. C. N. J. I. M. Nitin Hardeniya, "Natural Language Processing: Python and NLTK," oreilly, November 2016. [Online].

[47] S. Chakravarthy, "Simple Word Embedding for Natural Language Processing," towardsdatascience, 3 August 2020. [Online].

[48] N. Agarwal, "The Ultimate Guide To Different Word Embedding Techniques In NLP," kdnuggets, 4 November 2022. [Online].