Sentiment Analysis Based on Feature Selection and Machine Learning Techniques

Thajiba Tabassum

2015–2–60–022

Md Shohag

2016–1–60–047

Abdullahhil kafi

2016–1–60–076

Supervised by:

Lutfun Nahar Lota

Lecturer

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

Department of Computer Science and Engineering East West University

Dhaka-1212, Bangladesh

December, 2019

Declaration

We, hereby, declare that the work presented in this thesis is the outcome of the investigation performed by us under the supervision of Lutfun Nahar Lota, Lecturer, Department of Computer Science and Engineering, East West University. We also declare that no part of this thesis has been submitted elsewhere for the award of any degree or diploma.

Countersigned

. . . .

(**Lutfun Nahar Lota**)

Supervisor

Signature

. . . .

Thajiba Tabassum

(2015-2-60-022)

Signature

. . . .

Md Shohag

(2016-1-60-047)

Signature

. . . .

Abdullahhil kafi

(2016-1-60-076)

Signature

. . . .

Letter of Acceptance

This thesis report entitled “Sentiment Analysis Based on Feature Selection and Machine Learning Techniques” submitted by Thajiba Tabassum (2015-2-60-022), Md Shohag (2016-1-60-047) and Abdullahhil Kafi (2016-1-60-076) to the Department of Computer Science and Engineering, East West University is accepted by the department in partial fulfillment of requirements for the Award of the Degree of Bachelor of Science and Engineering in December, 2019.

Supervisor

……………………………………………………

(Lutfun Nahar Lota)
Lecturer
Department of Computer Science and Engineering
East WEST University, Dhaka

Chairperson

……………………………………………………..

(Taskeed Jabid)
Chairperson and Associate Professor,
Department of Computer Science and Engineering
East West University, Dhaka

Abstract

This research focuses on sentiment analysis of amazon food review dataset. For this work, it used some machine learning algorithms but specifically used sentiment analysis and LSTM. It was implemented by some machine learning algorithm with sentiment analysis. People are now a day more depends on restaurant food because of their busy life. For this they taste many kinds of restaurants and without knowledge they don't know where they can get good food with good service. They don't know which is good or which is bad and sometimes the reviews are so confusing that they do not understand. To overcome the above problems researchers have used machine learning algorithms to classify positive or negative food review with binary classification. In this study it discussed about long short term memory (LSTM) and four machine learning algorithm to solve this problem it was also use aspect based sentiment analysis. There are a lot of approaches developed for binary classification but it used Naïve Bayes (Bernoulli), Perceptron, Decision tree, Logistic regression, Long short term memory(LSTM). Recurrent Neural Networks(RNN) have exposed as widely used architectures and are united with sequence-based models.

The main research aims to develop this machine learning algorithm to gives the best result for binary classification in restaurant food review. The purpose of this research is binary classification and sentiment analysis. To classify text, it has been used Amazon food review data set. It has been used some machine learning algorithms. It has been divided the work into three stages, these were data preprocessing, sentiment analysis and binary classification.

Acknowledgments

For better understanding, a sample Acknowledgment is given below.

As it is true for everyone, we have also arrived at this point of achieving a goal in our life through various interactions with and help from other people. However, written words are often elusive and harbor diverse interpretations even in one's mother language. Therefore, we would not like to make efforts to find the best words to express my thankfulness other than simply listing those people who have contributed to this thesis itself in an essential way. This work was carried out in the Department of Computer Science and Engineering at East West University, Bangladesh.

First of all, we would like to express my deepest gratitude to the Almighty for His blessings on us. Next, our special thanks go to our supervisor, "Lutfun Nahar Lota", who gave us this opportunity, initiated us into the field of "Sentiment Analysis Based on Feature Selection and Machine Learning Techniques", and without whom this work would not have been possible. Her encouragement, visionaries and thoughtful comments and suggestions, unforgettable support at every stage of our B.Sc. study were simply appreciating and essential. Her ability to muddle us enough to finally answer our own question correctly is something valuable what we have learned and we would try to emulate, if ever we get the opportunity. we would like to thank our parents for their unending support, encouragement and prayers.

There are numerous other people too who have shown me their constant support and friendship in various
ways, directly or indirectly related to our academic life. We will remember them in our hearts and hope to
find a more appropriate place to acknowledge them in the future.

Thajiba Tabassum

December, 2019

Md Shohag

December, 2019

Abdullahhil Kafi

December, 2019

Table of Contents

Chapter 1

List of Figures

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Sentiment analysis is a natural language processing algorithm that identify the positive, neutral, and negative sentiment of texts. It is also known as Opinion Mining the field within Natural Language Processing (NLP) that builds systems that try to identify and extract opinions within text. Usually, besides identifying the opinion, these systems extract attributes of the expression. Currently, sentiment analysis is a topic of great interest and development. With the help of sentiment analysis systems, this unstructured information could be automatically transformed into structured data of public opinions about products, services, brands, politics, or any topic that people can express opinions about. This data can be very useful for commercial applications like marketing analysis, public relations, product reviews, net promoter scoring, product feedback, and customer service. There are a lot of approaches developed for binary classification but it used Naïve Bayes (Bernoulli), Perceptron, Decision tree, Logistic regression, Long short term memory(LSTM) and feature selection. Feature Selection is the process where you automatically or manually select those features which contribute most prediction variable or output in which were interested. Recurrent Neural Networks(RNN) have exposed as widely used architectures and were united with sequence-based models.

1. Research Objective

1.1 AIM

The research aims to develop some machine learning algorithm which gives the best result for binary classification in restaurant food review.

1.2 OBJECTIVE

- To implement some Machine learning algorithm on LSTM to gives the best result for binary classification in restaurant food review.

- To identify some Machine learning algorithm to gives the best result for binary classification.

.

1.3 PROBLEM STATEMENT

People are now a day more depends on restaurant food because of their busy life. For this they taste many kinds of restaurants and without knowledge they don't know where they can get good food with good service. They don't know which is good or which is bad and sometimes the reviews are so confusing that they do not understand. To overcome above problems researchers have used machine learning algorithms to classify positive or negative food review. In this study it discussed about long short term memory (LSTM) and four machine learning algorithm to solve this problem it was also use feature selection and sentiment analysis.

1.4 MOTIVATION

Now a day's people are so busy and they prefer restaurant food more than home cooked food, that's why peoples are much dependable on restaurant food. Day by day its increased. This is the advantages of startup restaurant but all restaurant does not have good service or good food or good rate. Then this restaurant review will help the customer to find out the best restaurant by their choice. It is personal choice to post a review about a restaurant, but sometimes it is beneficial for customers to know how costly it is, it can probably increase knowledge and judgement about restaurants, which can be shared with others, it's not just sharing info, it's also what comes to use when you are in different situations with different kinds of people. Sometimes the reviews are so confusing that it cannot be recognized with positive or negative review then this sentiment analysis on restaurant food review can help them out to get proper review.

1.4 STRUCTURE

It is vital for a research to be well organized. For better understanding, it organized the remaining portion of this paper within the taking after sequences.

In chapter 2: It was talked about almost background work. Where include a few extra data related to this research.

In chapter 3: It was talked about the related paper works of binary classification and sentiment analysis. Many researchers were already worked with this topic. There was a discussion about sentiment analysis and binary classification.

In chapter 4: In this section, it proposed methodology where the discuss about the mechanism of this research work process.

In chapter 5: It discussed the result analysis. It has been mainly focused on combination of different machine learning algorithm and recurrent neural network (LSTM).

In chapter 6: It discussed about conclusion and Future work.

Chapter 2

Background

This research is based on binary classification of restaurant food review and it worked with amazon food review dataset. For this work, it used some machine learning algorithms but specifically used feature selection, sentiment analysis and LSTM. It was implemented by some machine learning algorithm with sentiment analysis. There are a lot of approaches developed for binary classification but it used Naïve Bayes (Bernoulli), Perceptron, Decision tree, Logistic regression, Long short term memory (LSTM).

2.1 Binary Classification

With binary classification the dataset was classified with two groups like positive and negative. The text or comment was negative or positive, binary classification classified it. In this research study if anyone posted any kind of review about restaurant food then with binary classification it was classified with positive and negative group. And if anyone rate the food then it calculates with If data rating was more than 3 then it indicated positivity and below 3 it indicated negativity.

2.2 Sentiment Analysis

The most excellent businesses get it the opinion of their customers—what individuals are saying, how they're saying it, and what they cruel. Client opinion can be found in tweets, comments, surveys, or other places where individuals specify the brand. Opinion Examination is the space of understanding these feelings with program, and it's a must-understand for engineers and trade pioneers in an advanced workplace. As with numerous other areas, progresses in deep learning have brought sentiment analysis into the frontal area of cutting-edge calculations. It utilizes natural language processing, statistics, and text analysis to extract, and identify the sentiment of text into positive, negative, or unbiased categories.

Sentiment analysis can offer assistance to naturally change the unstructured data into organized information of open conclusions approximately items, administrations, brands, legislative issues or any other theme that individuals can express conclusions approximately. This information can be exceptionally valuable for commercial applications like showcasing investigation, open relations, item surveys, net promoter scoring, item input, and client benefit.

2.3 Feature Selection

Feature selection is one of the major concepts in machine learning which was hugely impacts the performance of this study. Some time it was difficult to identify the related feature from a set of data and removing the irrelevant or less important feature with that the study did not get better accuracy. Feature selection train the data and help to achieve desired result. Feature Selection is the process where you automatically or manually select those features which contribute most prediction variable or output in which were interested. It reduces overfitting, training time and improve accuracy.

2.4 LSTM

It cleans noise from dataset then choose the most frequent word and train the model and predict the positivity and negativity. In this study it trained about 70000 data and test 30000 data. It handled the issue of long-term conditions of RNN in which the RNN cannot predict the word put away within the long term memory but can donate more exact forecasts from the later data. As the crevice length increments RNN does not grant efficient execution. LSTM can by default hold the data for long period of time. It is utilized for handling, anticipating and classifying on the premise of time arrangement information.

Chapter 3

Related work

The National Investigate Chamber of Canada [5] connected Multi Course Bolster Vector Machine (SVM) and lexicon based approach as an extra highlight within the classification process. It's classification execution within the shape of F1 Degree is 88.57%. In expansion, investigate conducted by the Xerox Inquire about Middle Europe [6] utilizing typical parser outlined with extraordinary vocabulary and combined with SVM gotten F1-Measure of 82.28%. Another ponder by the College of West Bohemia [7] utilizing Most Extreme Entropy classifier with 12 highlights such as words, LDA, bigrams, word clusters, tf-idf, and other highlights given F1-Measure of 81.04%.

An investigate conducted by Cities [8] proved that Naïve Bayes gave long exhibitions within the prepare of estimation examination on English Tweets. A comparable consider conducted by Prasad [9] appeared that Naïve Bayes was able to grant long execution within the prepare of assumption investigation on micro-blogging. In expansion, the inquire about on assumption investigation conducted by Xhemali et al. in which comparing three strategies (e.g., Naïve Bayes, Choice Tree, and Neural Systems) appeared that Naïve Bayes was extraordinary against the two other strategies. Be that as it may, the three said inquires about connected Naïve Bayes classifiers for classifying assumption extremity as it were, and the classification of estimation extremity was done at sentence level not at angle level. Whereas, this inquire about proposed the utilization of Naïve Bayes to distinguish a perspective on item audits additionally to classify the opinion extremity of the viewpoint - i.e., estimation examination on angle level. Naïve Bayes is combined with Chi Square strategy as well as POS labeling as an include choice.

It executed distinctive models to illuminate the audit value classification issue. Both feed-forward neural arrange and LSTM were able to defeat the pattern show. Exhibitions of the models were assessed utilizing 0-1 misfortune and F-1 scores. In common, LSTM beat feed-forward neural arrange, as it prepared the possess word vectors in that demonstrate, and LSTM itself was able to store more data because it forms arrangement of words. Other than, it built a recommender framework utilizing the user-item-rating information to assist examine this dataset and expecting to create association with audit classification. The execution of recommender framework is measured by RMSE in rating expectations. [2]

With the rise of online shopping and social systems, audits and appraisals has ended up a vital way to get it the assumption and require of clients. Being able to naturally channel the surveys has gotten to be a key challenge for those companies and websites. As a result, a parcel of exertion has been made to NLP assignments such as wistful examination [1].

As the computing control increments, neural systems have gotten to be increasingly prevalent for dialect modeling. Both feed-forward neural systems [2] and repetitive neural systems have been utilized as often as possible in managing with content categorization, archive labeling and consecutive word producing issues. The LSTM (long short-term memory) neural arrange [3] is also gaining increasingly consideration in their execution in discourse recognition.

A parcel of past work has been drained building recommender frameworks as well. Collaborative sifting [4], Content-based sifting [5] and Matrix Factorization [6] have been the foremost fruitful ones with wide application in industry. As of late, profound Learning has been used for Music Suggestion [7].

Repetitive Neural Arrange is additionally received for collaborative sifting-based proposal as well.

Audits portray assumptions of clients towards different viewpoints of an item or benefit. A few of these angles can be assembled into coarser viewpoint categories. SemEval-2014 had a shared assignment (Errand 4) on aspect-level opinion examination, with over 30 groups taken an interest. In this paper, it portrays the entries, which stood to begin with in identifying perspective categories, to begin with in identifying assumption towards perspective categories, third in identifying viewpoint terms, and to begin with and moment in identifying opinion towards angle terms within the portable workstation and eatery spaces, respectively. [3]

In this venture, it thinks about the applications of Recursive Neural Organize on estimation examination errands. To handle the crude content information from Amazon Fine Nourishment Surveys, it proposes and actualize a method to parse parallel trees utilizing Stanford NLP Parser. In expansion, moreover propose a novel method to name tree nodes in arrange to realize the level of supervision that RNN requires, within the setting of the need of labeling within the original dataset. At long last, it proposes an unused demonstrate RNNMS (Recursive Neural Organize for Different Sentences), and have superior comes about than the pattern in terms of each metric it considers. [4]

Recent work has been centered on other complicated RNN models such as recursive neural tensor organizes [1], which is strong in recognizing refuting negatives, and Tree LSTM [2], which has the thought of disregard door acquired from LSTM. may be a hot demonstrate and certainly worth the ponders in an extend? Looking at final years extend [3], the exactness of that was 59.32% to 63.71%, depending on distinctive Recursive Neural Network models. It created vanilla one covered up layer, two covered up layer recursive neural systems and RNTN. In venture, it achieved 10% more than the result, which may be a noteworthy change. Superior tree parser and open up labeling inside hubs strategies are credited to the way better result.
In the meantime, Stanford Treebank, due to the solid supervision, that's to say, altogether labeled inner hubs, accomplished exceptionally great test precision (more than 80%). It is so distant the leading information set which to be used for Recursive Neural Network. On the other hand, it will expect that lack of labeling is one of the enormous challenges for Recursive Neural Network.
Back to the Kaggle challenge, in spite of the fact that there's no current champ exactness right now, the information set and the questions were really drawn from a paper coming from Stanford. [4] In spite of the fact that, in their paper, the most challenge was not estimation examination, the most noteworthy test exactness was around 40% in considers of clients tastes and inclinations changing and advancing over time. This moo exactness too appeared that this was a challenging information set to analyze on.

Online Client audits may be incredible stage for collecting huge volumes of data for assumption investigation. Here it proposes a framework which classifies the client audits into two fundamental categories: Positive and Negative audits. The Classification calculation employments as it were the in general survey scores to get it the opinion behind each survey and extricate the critical perspectives almost the item. It created a proficient classifier show to classify the given audit is either a positive survey or negative survey by analyzing the execution of different classification calculations on the audit information corpus. Clustering strategies are then used to recognize key opinion characteristics to supply them to the clients, which makes a difference the client to get it the angles of the products/services they wish to purchase or involvement. [5]

Supervised learning is additionally a common approach for angle extraction. Hercig et al. [4] utilized CRF with lexical and syntactic highlights, such as a token, POS tag, reliance way, and highlights from Persistent Bag-of-Word (CBOW), which is one of distributional semantic models [7]. Distributional semantic demonstrate is based on the suspicion that the meaning of a word can be gathered from it utilizations [8].

Lexicon-based approach [9] and SentiWordNet-based approach [10] can be utilized for assumption classification. Hercig et al. [4] utilized administered learning to classify estimation of categories in a sentence. It utilizes Most Extreme Entropy (MaxEnt) classifier [11] with lexical highlights, syntactic highlights, and highlights from CBOW and Worldwide Vectors for word representation (GloVe) [12].

Gojali and Khodra [5] blended viewpoint and assumption extraction into one step by utilizing CRF with token and POS tag as highlights. Each token is relegated into one of four substances: angle, positive conclusion, negative supposition, and other. After naming each token, angle is combined by its coordinating supposition. At that point, the introduction or extremity of the supposition words are decided by checking the nearness of invalidation word with five words separate from the supposition words.

After angle extraction, the viewpoints can be categorized into categories. One approach is to utilize WordNet. Liu et al. [13] utilize WordNet to discover equivalent word between perspective. For illustration, photo, picture, and picture can be categorized into one category since those three words are equivalent words in WordNet. Gojali and Khodra [5] moreover utilize WordNet to discover likeness between perspective and predefined seed word for eatery categories. Hercig et al. [4] utilize administered learning utilizing multilabel classification with MaxEnt calculation. It too utilizes lexical and syntactic highlights as in perspective extraction and assumption classification step and highlights from CBOW, GloVe, and Idle Dirichlet Assignment (LDA) [14].

With the approach of social media, the number of surveys for any specific item is in millions, as there exist thousands of websites where that specific item exists. As the numbers of audits are exceptionally tall the client closes up investing a part of time looking for the finest item based on the encounters shared by survey journalists. This paper presents an opinion based rating approach for nourishment formulas which sorts nourishment formulas show on different websites on the premise of assumptions of review writers. The comes about are appeared with the assistance of a versatile application: Foodoholic. The yield of the application is a requested list of formulas with client input as their center fixing. [7]

The development of web contributes a gigantic amount of client made substance such as client criticism, suppositions and surveys. Opinion investigation in web grasps the issue of amassing information within the web and extraction around conclusions. Examining the suppositions of clients makes a difference to decide the people's sentiments almost an item and how it is gotten within the showcase. Different commercial apparatuses are accessible for assumption examination. In this paper, it proposes a framework which classifies the surveys on a scale of 1 to 5 based on the feelings within the words. The bunches of words utilized to form a choice to rate the surveys are shown as word cloud. [8]

Sentiment investigation strategies are machine learning, vocabulary based and crossover strategies. Machine learning methods are implemented in directed classification. Vocabulary based approach depends on the collection of supposition terms. Combined approach of vocabulary and machine learning may be a cross breed approach. The computational methods including dialects are called normal dialect preparing.

Most of the existing investigates center on the mining of information accessible online in web. Sometime recently the World Wide Web, data or conclusion almost the items is collected based on studies physically. Xing et al. [8] had proposed a work on item audits collected from amazon to recognize the refutation expressions. Sentence level and audit level classification of information is performed for the information collected from February to April 2014. Aashutosh Bhatt et al. [3] utilized audits of iPhone 5 extricated from Amazon site and proposed a run the show based extraction of item include assumption examination. POS method is executed to each and each sentence level and the comes about are appeared in charts. Ahmad Kamal [2] utilized directed and run the show based strategies to mine the suppositions from online item audits.

Chapter 4

Methodology

The purpose of this research is binary classification and sentiment analysis. To classify text, it has been used Amazon food review data set. It has been used some machine learning algorithms. It has been divided the work into three stages, these were data preprocessing, sentiment analysis and binary classification. A discussion on this approach is described below: --

4.1 Data preprocessing

Data preprocessing is a data mining technique that convert raw data into clean data set. For achieved better results it need to some specified Machine Learning model needs information in a specified format. The main aspect was that data set has been formatted in such Machine Learning algorithm like naïve Bayes (Bernoulli), logistic regression, perceptron and decision tree. Also Deep Learning algorithms like LSTM were executed. If data rating was more than 3 then it indicated positivity and below 3 it indicated negativity. There was some significant amount of reviews, That was not classified positivity or negativity.

4.2 Sentiment Analysis

Sentiment analysis finds and legitimizes the opinion of the individual with regard to a given source of substance. Social media contain colossal sum of the assumption information within the shape of tweets, blogs, and upgrades on the status, posts, etc. finds and legitimizes the opinion of the individual with regard to a given source of substance. Social media contain colossal sum of the assumption information within the shape of tweets, blogs, and upgrades on the status, posts, etc. If the speaker communicates a positive or negative supposition, the thing that's being talked around, the individual, or substance that communicates the opinion.

4.3 Binary Classification

Binary or binomial classification is the task of classifying the elements of a given set into two groups on the basis of a classification rule. In this research, it has been classified an Amazon food review data set in to two categories, positivity and negativity by using some machine learning algorithms. Machine Learning algorithm which were used in this research described below: -

4.4 Feature Selection

Feature selection is one of the major concepts in machine learning which was hugely impacts the performance of this study. Some time it was difficult to identify the related feature from a set of data and removing the irrelevant or less important feature with that the study did not get better accuracy. Feature selection train the data and help to achieve desired result. Feature Selection is the process where you automatically or manually select those features which contribute most prediction variable or output in which were interested. It reduces overfitting, training time and improve accuracy.

Algorithms

4.5 Naïve Bayes (Bernoulli)

Naïve Bayes algorithm is used for text classification in machine learning. There is different version of naïve Bayes like Bernoulli.

Bernoulli

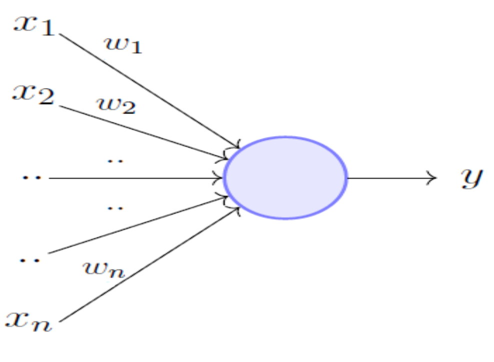
Bernoulli deals with Boolean value. It generates 0s or 1s for predicting. If each term of the vocabulary equal to 1, the term belongs to examining documents. If that is 0, the term does not belong to examining documents. Non-accruing terms in document are takes into document and they are factored when computing the conditional probabilities and thus the absence of term is taken into account. It assumes that all our features are binary such that they take only two values, 0s can represent word absence in the document "and 1s as "word present in the document"

4.6 Logistic Regression

Logistic Regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be class variable, i.e. 0-no, 1- yes. Therefore, we are squashing the output of the linear equation into a range of [0,1]. To squash the predicted value between 0 and 1, we use this function.

4.7 Perceptron

A single perceptron can only be used to implement linearly separable functions. It takes both real and Boolean inputs and associates a set of weights to them, along with a bias (the threshold thing I mentioned above). We learn the weights; we get the function. A perceptron is not the Sigmoid neuron we use in ANNs or any deep learning networks today. It takes an input, aggregates it (weighted sum) and returns 1 only if the aggregated sum is more than some threshold else returns 0.



$$y = 1 \quad if \sum_{i=1}^{n} w_i * x_i \geq \theta$$

$$= 0 \quad if \sum_{i=1}^{n} w_i * x_i < \theta$$

Rewriting the above,

$$y = 1 \quad if \sum_{i=1}^{n} w_i * x_i - \theta \geq 0$$

$$= 0 \quad if \sum_{i=1}^{n} w_i * x_i - \theta < 0$$

4.8 Decision Tree

The understanding level of Decision Trees algorithm is so easy compared with other classification algorithms. The decision tree algorithm tries to solve the problem, by using tree representation. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label. In decision trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

4.9 LSTM (Long Short-Term Memory)

Sequence prediction problems have been around for a long time. With the recent breakthroughs that have been happening in data science, it is found that for almost all of these sequence prediction problems, Long Short Term Memory networks, LSTMs have been observed as the most effective solution. Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Long Short Term Memory is a kind of recurrent neural network. It cleans noise from dataset then choose the most frequent word and train the model and predict the positivity and negativity. In this study it trained about 70000 data and test 30000 data. In RNN output from the last step is fed as input in the current step. In RNN yield from the final step is bolstered as input within the current step. It handled the issue of long-term conditions of RNN in which the RNN cannot predict the word put away within the long term memory but can donate more exact forecasts from the later data. As the crevice length increments RNN does not grant efficient execution. LSTM can by default hold the data for long period of time. It is utilized for handling, anticipating and classifying on the premise of time arrangement information.

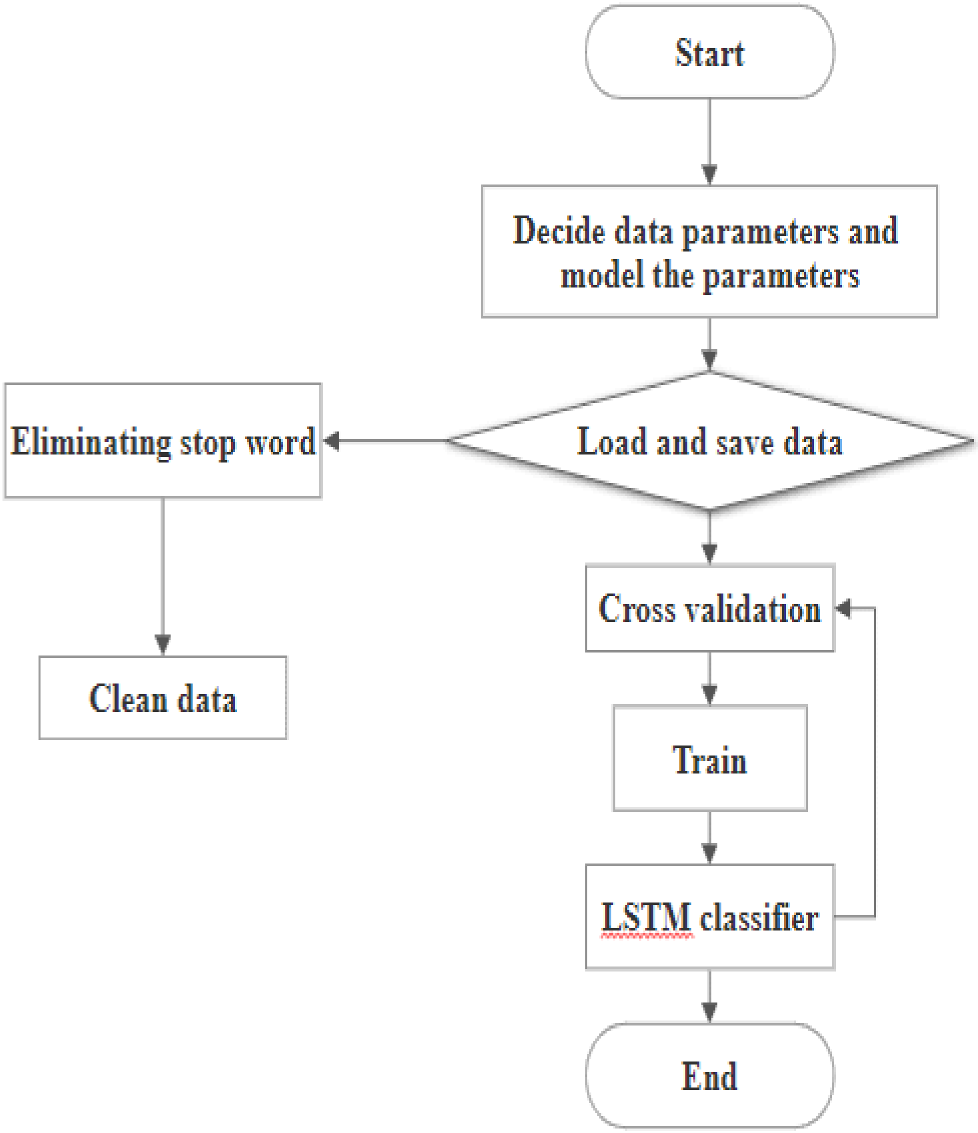


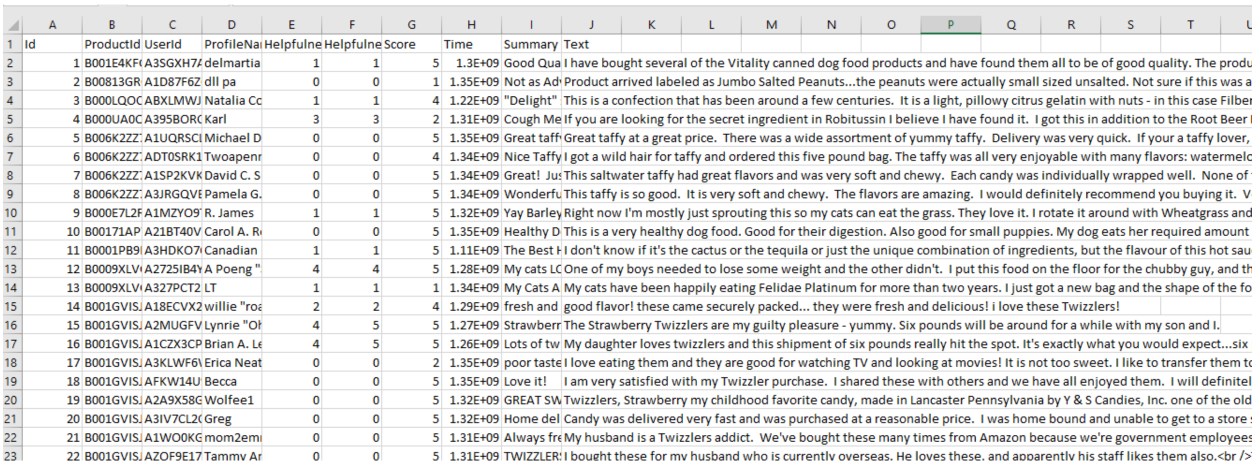
Fig: 01 LSTM Procedure

Chapter 5

Result

5.1 Data set

This study used amazon food review dataset for binary classification. In this review dataset consists of total 568454 data. Here is the sample dataset



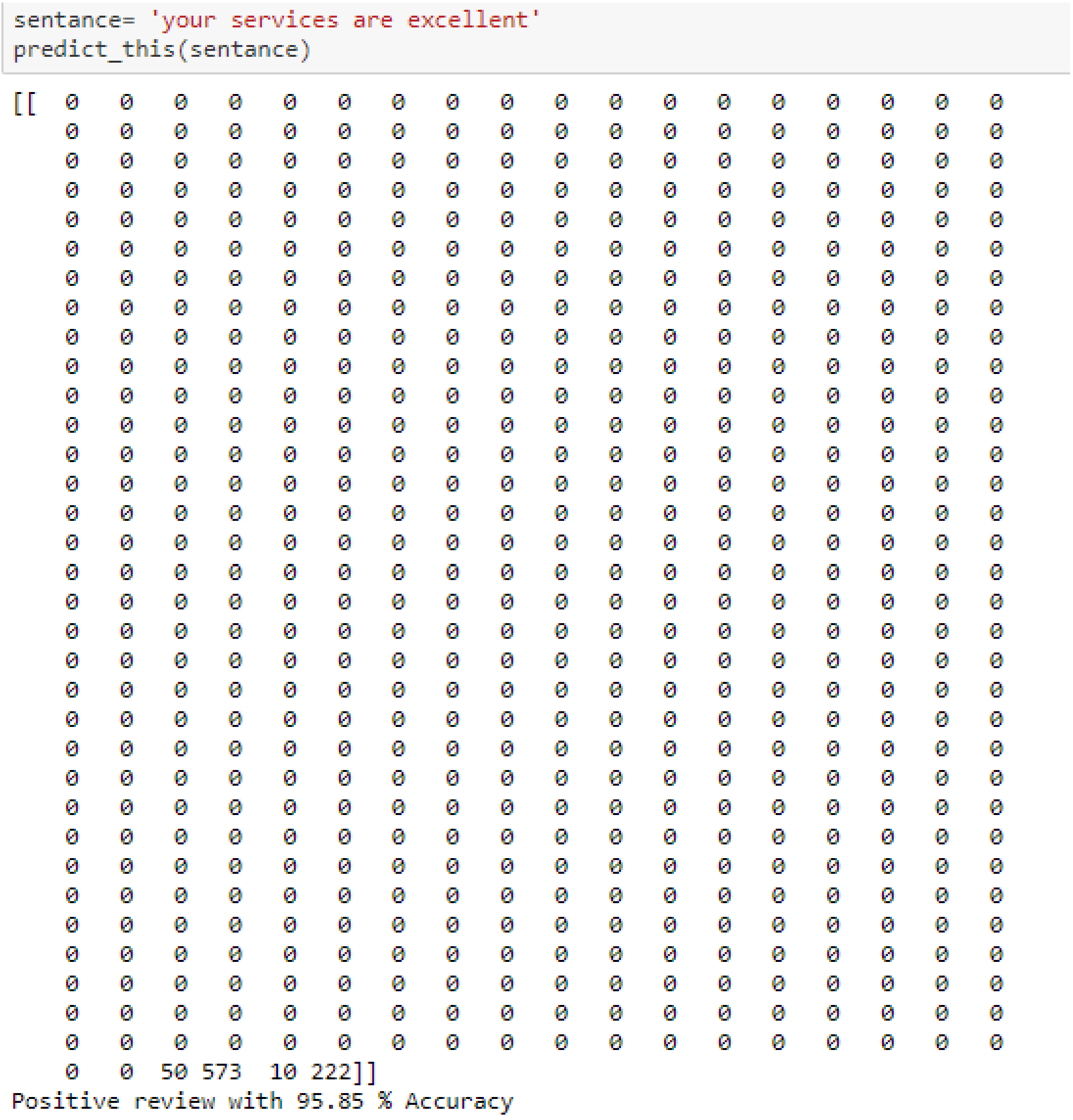
	A	B	C	D	E	F	G	H	I	J
1	Id	ProductId	UserId	ProfileNa	Helpfulne	Helpfulne	Score	Time	Summary	Text
2	1	B001E4KF	A3SGXH7A	delmartia	1	1	5	1.3E+09	Good Qua	I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The produ
3	2	B00813GR	A1D87F6Z	dll pa	0	0	1	1.35E+09	Not as Ad	Product arrived labeled as Jumbo Salted Peanuts...the peanuts were actually small sized unsalted. Not sure if this was a
4	3	B000LQOC	ABXLMWJ	Natalia Cc	1	1	4	1.22E+09	"Delight"	This is a confection that has been around a few centuries. It is a light, pillowy citrus gelatin with nuts - in this case Filber
5	4	B000UA0C	A395BORC	Karl	3	3	2	1.31E+09	Cough Me	If you are looking for the secret ingredient in Robitussin I believe I have found it. I got this in addition to the Root Beer
6	5	B006K2ZZ	A1UQRSCI	Michael D	0	0	5	1.35E+09	Great taff	Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover,
7	6	B006K2ZZ	ADT0SRK1	Twoapenr	0	0	4	1.34E+09	Nice Taffy	I got a wild hair for taffy and ordered this five pound bag. The taffy was all very enjoyable with many flavors: watermelc
8	7	B006K2ZZ	A1SP2KVK	David C. S	0	0	5	1.34E+09	Great! Jus	This saltwater taffy had great flavors and was very soft and chewy. Each candy was individually wrapped well. None of
9	8	B006K2ZZ	A3JRGQVE	Pamela G.	0	0	5	1.34E+09	Wonderfu	This taffy is so good. It is very soft and chewy. The flavors are amazing. I would definitely recommend you buying it. V
10	9	B000E7L2F	A1MZYO9	R. James	1	1	5	1.32E+09	Yay Barley	Right now I'm mostly just sprouting this so my cats can eat the grass. They love it. I rotate it around with Wheatgrass and
11	10	B00171AP	A21BT40V	Carol A. R	0	0	5	1.35E+09	Healthy D	This is a very healthy dog food. Good for their digestion. Also good for small puppies. My dog eats her required amount
12	11	B0001PB9	A3HDKO7	Canadian	1	1	5	1.11E+09	The Best H	I don't know if it's the cactus or the tequila or just the unique combination of ingredients, but the flavour of this hot sau
13	12	B0009XLV	A2725IB4Y	A Poeng "	4	4	5	1.28E+09	My cats LC	One of my boys needed to lose some weight and the other didn't. I put this food on the floor for the chubby guy, and th
14	13	B0009XLV	A327PCT2	LT	1	1	1	1.34E+09	My Cats A	My cats have been happily eating Felidae Platinum for more than two years. I just got a new bag and the shape of the fo
15	14	B001GVISJ	A18ECVX2	willie "roa	2	2	4	1.29E+09	fresh and	good flavor! these came securely packed... they were fresh and delicious! i love these Twizzlers!
16	15	B001GVISJ	A2MUGFV	Lynrie "Oh	4	5	5	1.27E+09	Strawberr	The Strawberry Twizzlers are my guilty pleasure - yummy. Six pounds will be around for a while with my son and I.
17	16	B001GVISJ	A1CZX3CP	Brian A. Le	4	5	5	1.26E+09	Lots of tw	My daughter loves twizzlers and this shipment of six pounds really hit the spot. It's exactly what you would expect...six
18	17	B001GVISJ	A3KLWF6\	Erica Neat	0	0	2	1.35E+09	poor taste	I love eating them and they are good for watching TV and looking at movies! It is not too sweet. I like to transfer them tc
19	18	B001GVISJ	AFKW14U	Becca	0	0	5	1.35E+09	Love it!	I am very satisfied with my Twizzler purchase. I shared these with others and we have all enjoyed them. I will definitel
20	19	B001GVISJ	A2A9X58G	Wolfee1	0	0	5	1.32E+09	GREAT SW	Twizzlers, Strawberry my childhood favorite candy, made in Lancaster Pennsylvania by Y & S Candies, Inc. one of the old
21	20	B001GVISJ	A3IV7CL2	Greg	0	0	5	1.32E+09	Home del	Candy was delivered very fast and was purchased at a reasonable price. I was home bound and unable to get to a store
22	21	B001GVISJ	A1WO0KG	mom2em	0	0	5	1.31E+09	Always fre	My husband is a Twizzlers addict. We've bought these many times from Amazon because we're government employees
23	22	B001GVISJ	AZOF9E17	Tammy Ar	0	0	5	1.31E+09	TWIZZLER:	I bought these for my husband who is currently overseas. He loves these, and apparently his staff likes them also.<br />

Fig:02 shows the dataset of amazon food review.

5.2 Result Analysis

The result based on accuracy for basic machine learning algorithm is shown below. In this study, it used Kaggle kernel tool which is online based on tool. This study used amazon food review dataset for binary classification. And the dataset consists of total 568454 data. The result based on accuracy for basic machine learning algorithm. With preprocessing, Bernoulli showed 78%, Perceptron showed 84%, Logistic showed 85%, Decision tree showed 77%. And Without preprocessing Bernoulli showed 76%, Perceptron showed 81%, Logistic Regression showed 83%, Decision tree showed 74%. All this results were graphically showed for better understanding. It also showed that LSTM is the better performed. It applied LSTM on single reviews and its says some case positivity 95.85% and says the negative word for negativity 79.08%. Show it by snapshot.

LSTM Positivity Review Accuracy:

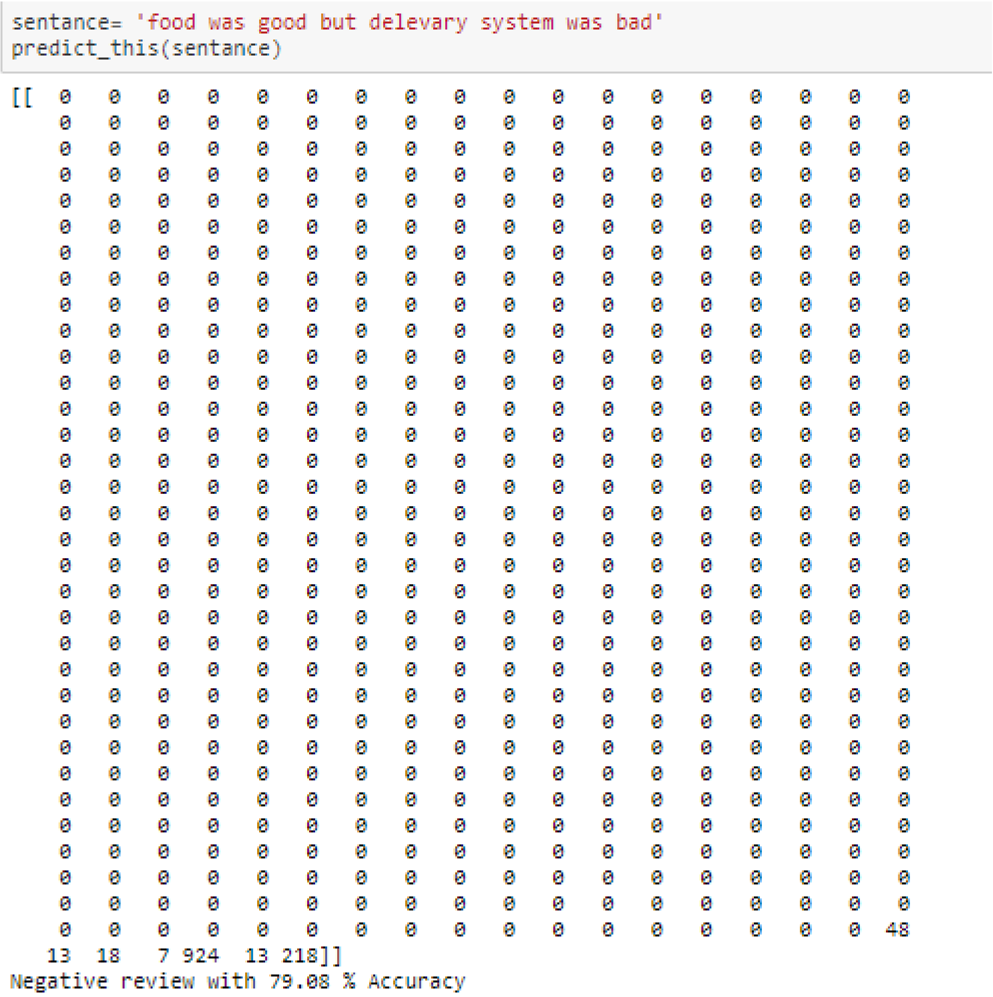


```
sentance= 'your services are excellent'
predict_this(sentance)
```

```
[[  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0  50 573  10 222]]
Positive review with 95.85 % Accuracy
```

Fig:03 shows the LSTM positivity review accuracy.

LSTM negativity Review Accuracy:



```
sentance= 'food was good but delevary system was bad'
predict_this(sentance)
```

```
[[  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  48
   13  18   7 924  13 218]]
Negative review with 79.08 % Accuracy
```

Fig:04 shows the LSTM negativity review accuracy.

Preprocessing Stage: In this stage all the machine learning algorithm perform better result.

Algorithm	Accuracy
Bernoulli	0.786
perceptron	0.84
Logistic Regression	0.85
Decision Tree	0.77

Table: 01 its shows that the Machine Learning algorithm and its Accuracy.

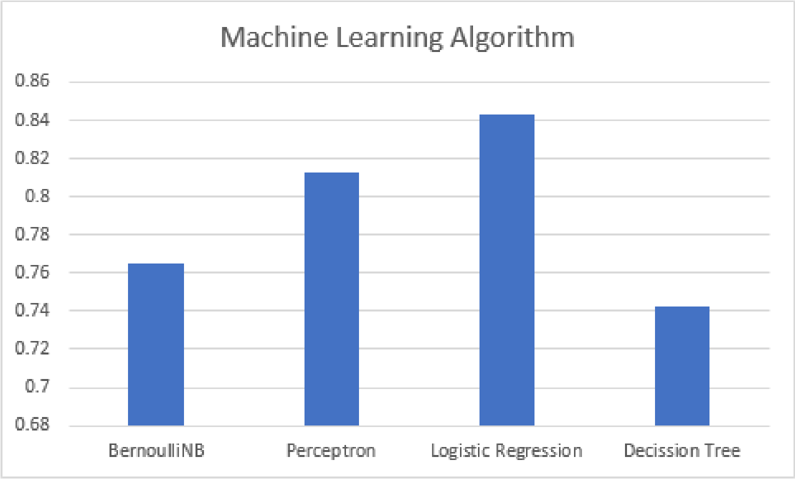
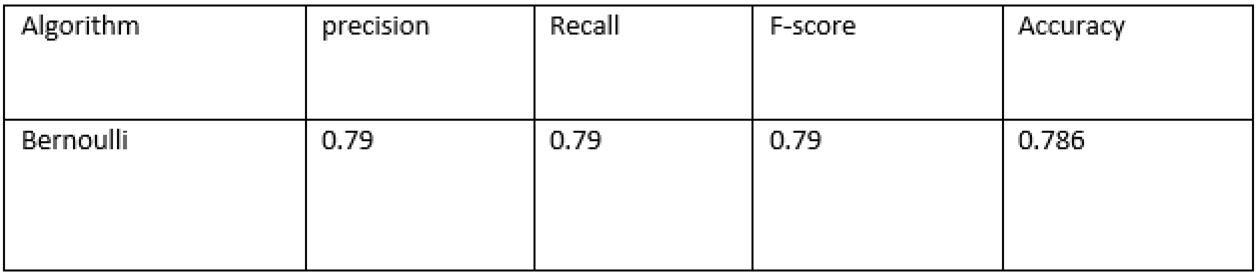


Fig: 05 shows that the Machine Learning algorithm and its Accuracy by the plot



Algorithm	precision	Recall	F-score	Accuracy
Bernoulli	0.79	0.79	0.79	0.786

Table: 02 shows that the Bernoulli algorithm and its precision, Recall, F-score and finally its accuracy

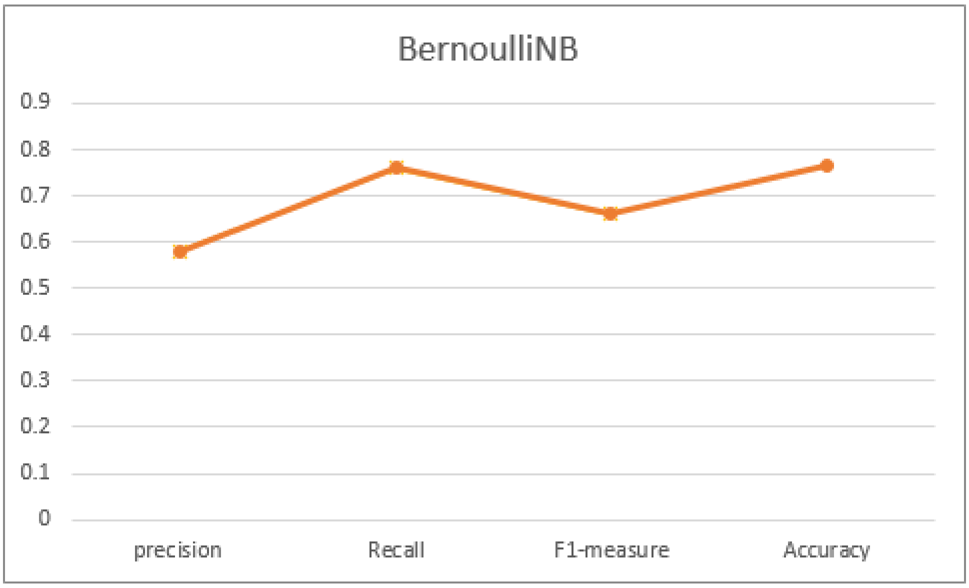
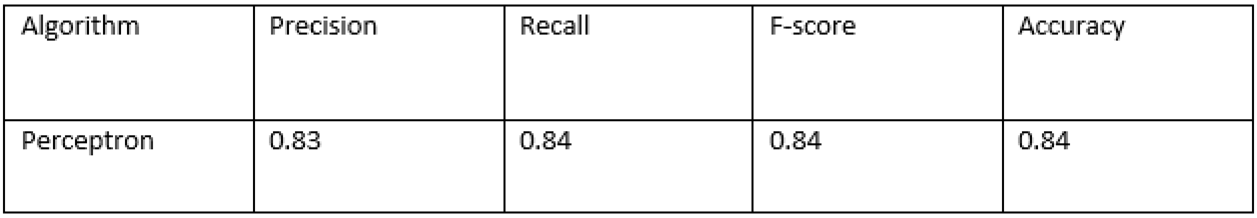


Fig: 06 shows that the actual plot of the Bernoulli algorithm for given table 02 values



Algorithm	Precision	Recall	F-score	Accuracy
Perceptron	0.83	0.84	0.84	0.84

Table: 03 shows that the Perceptron algorithm and its precision, Recall, F-score and finally its accuracy

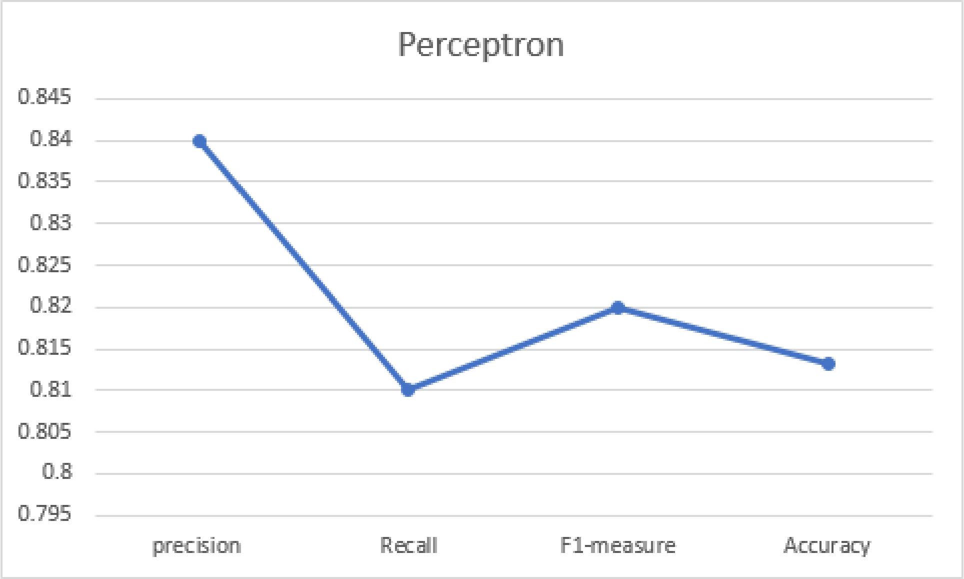
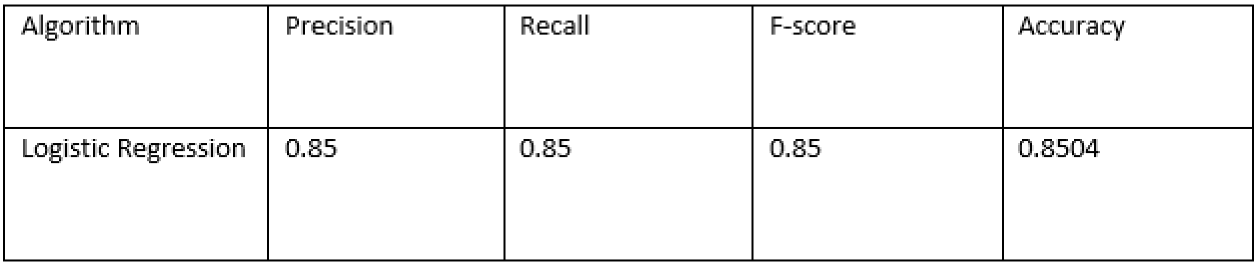


Fig: 07 shows that the actual plot of the Perceptron algorithm for given table 03 values



Algorithm	Precision	Recall	F-score	Accuracy
Logistic Regression	0.85	0.85	0.85	0.8504

Table: 04 shows that the Logistic Regression algorithm and its precision, Recall, F-score and finally its accuracy

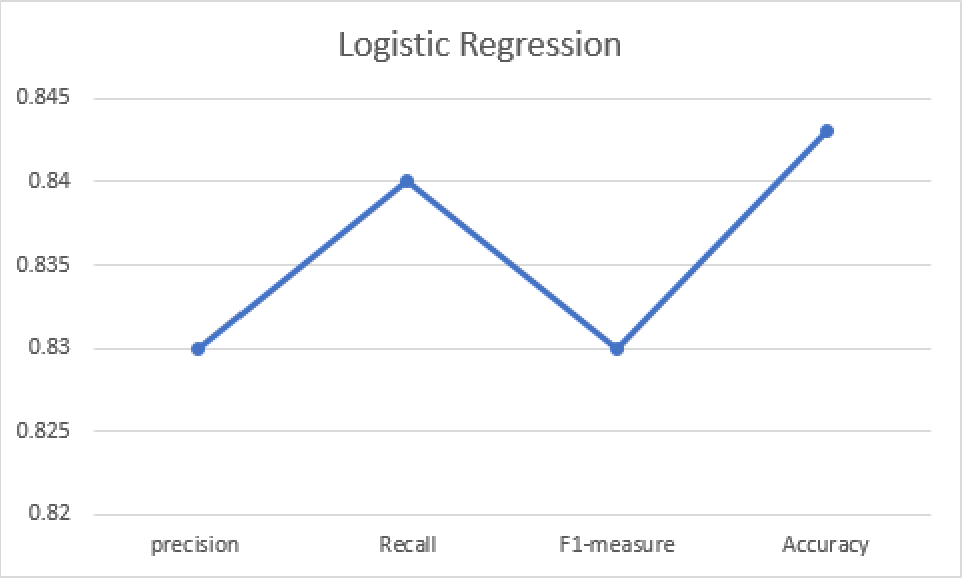


Fig: 08 shows that the actual plot of the Logistic Regression algorithm for given table 04 values

Algorithm	Precision	Recall	F-score	Accuracy
Decision Tree	0.77	0.78	0.77	0.777

Table: 05 shows that the Decision Tree algorithm and its precision, Recall, F-score and finally its accuracy

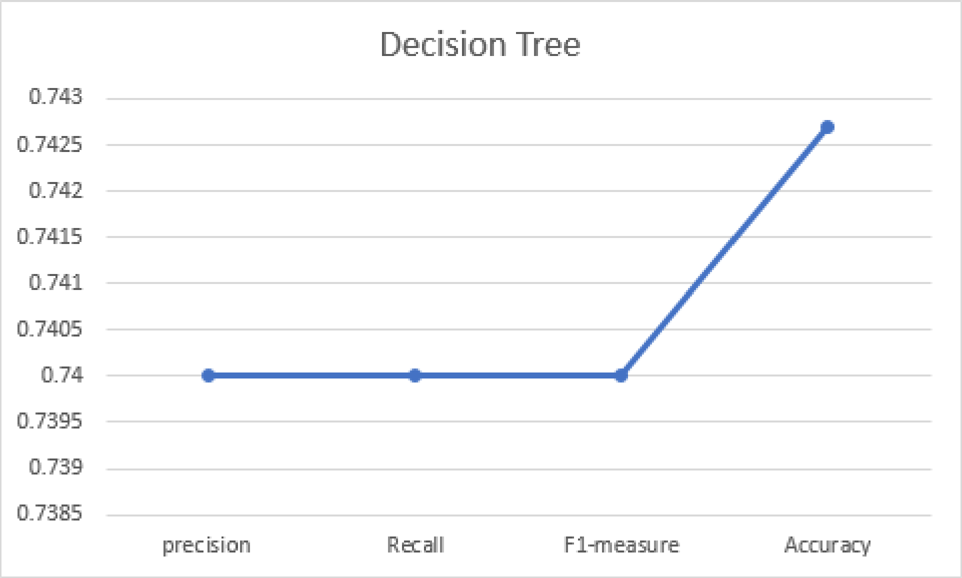


Fig: 09 shows that the actual plot of the Decision Tree algorithm for given table 05 values.

Without preprocessing Stage: This stage algorithm performed normal result and bad accuracy.

Algorithm	Accuracy
Bernoulli	0.764
Perceptron	0.816
Logistic Regression	0.837
Decision Tree	0.747

Table: 06 its shows that the Machine Learning algorithm and its Accuracy.

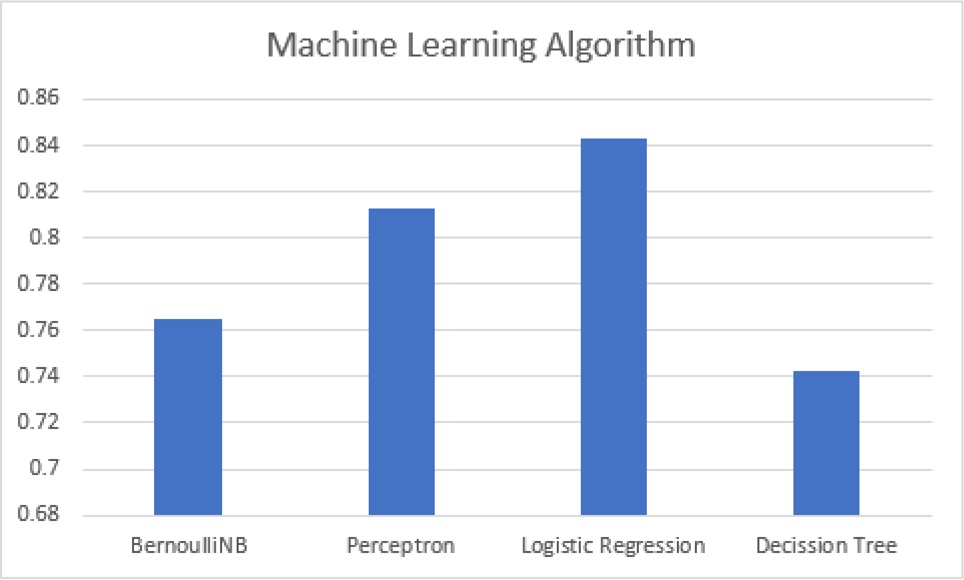


Fig: 10 shows that the Machine Learning algorithm and its Accuracy by the plot

Algorithm	precision	Recall	F-score	Accuracy
Bernoulli	0.58	0.76	0.66	0.764

Table: 07 shows that the Bernoulli algorithm and its precision, Recall, F-score and finally its accuracy

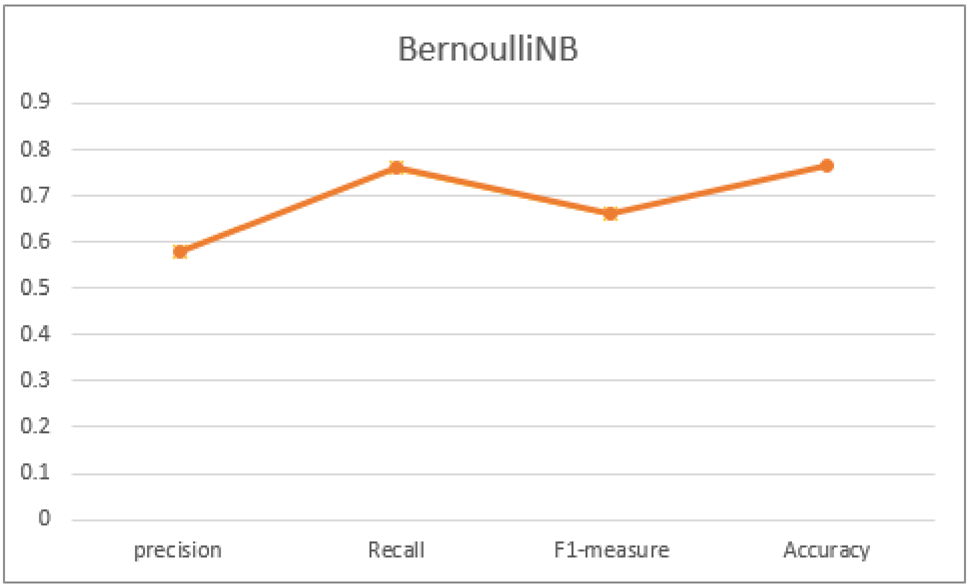


Fig: 11

Fig 11 shows that the actual plot of the Bernoulli algorithm for given table 07 values

Algorithm	Precision	Recall	F-score	Accuracy
Perceptron	0.82	0.82	0.82	0.816

Table: 08 shows that the perceptron algorithm and its precision, Recall, F-score and finally accuracy

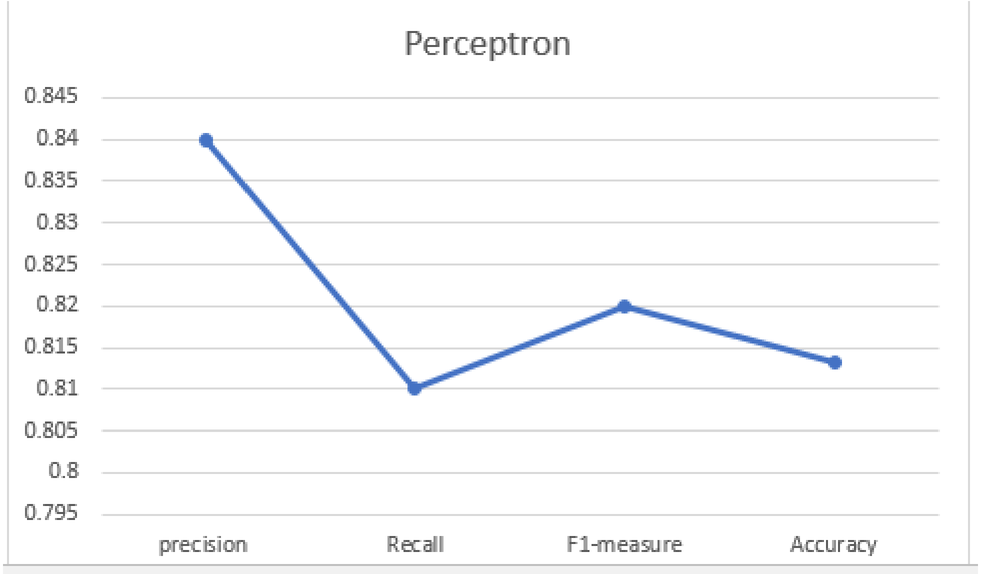


Fig: 12 shows that the actual plot of the perceptron algorithm for given table 08 values

Algorithm	Precision	Recall	F-score	Accuracy
Logistic Regression	0.83	0.84	0.83	0.837

Table: 09 shows that the Logistic Regression algorithm and its precision, Recall, F-score and finally its accuracy

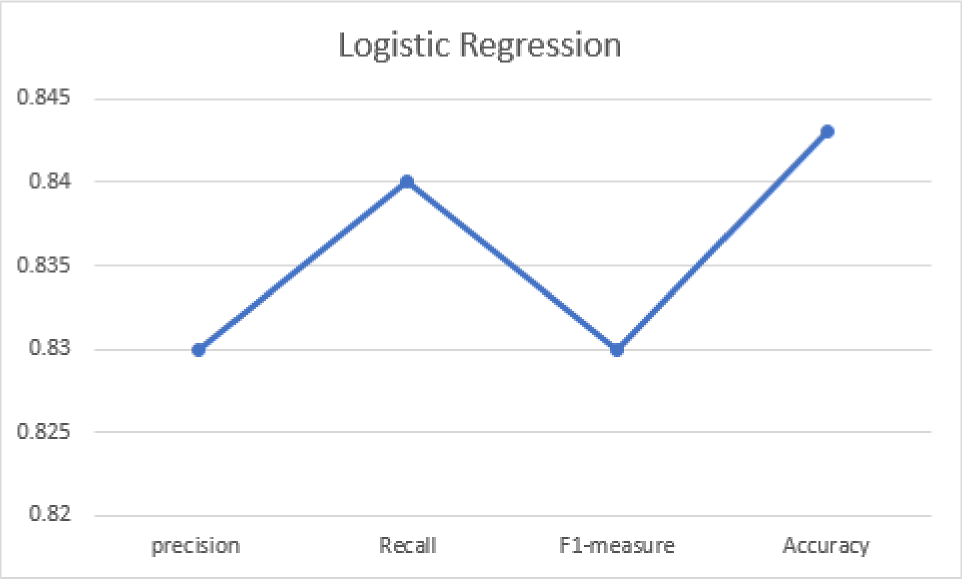
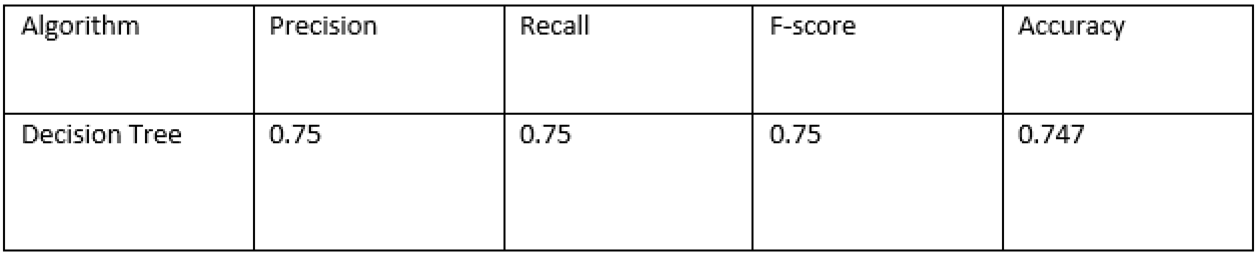


Fig: 13 shows that the actual plot of the Logistic Regression algorithm for given table 09 values



Algorithm	Precision	Recall	F-score	Accuracy
Decision Tree	0.75	0.75	0.75	0.747

Table: 10 shows that the Decision Tree algorithm and its precision, Recall, F-score and finally its

Accuracy.

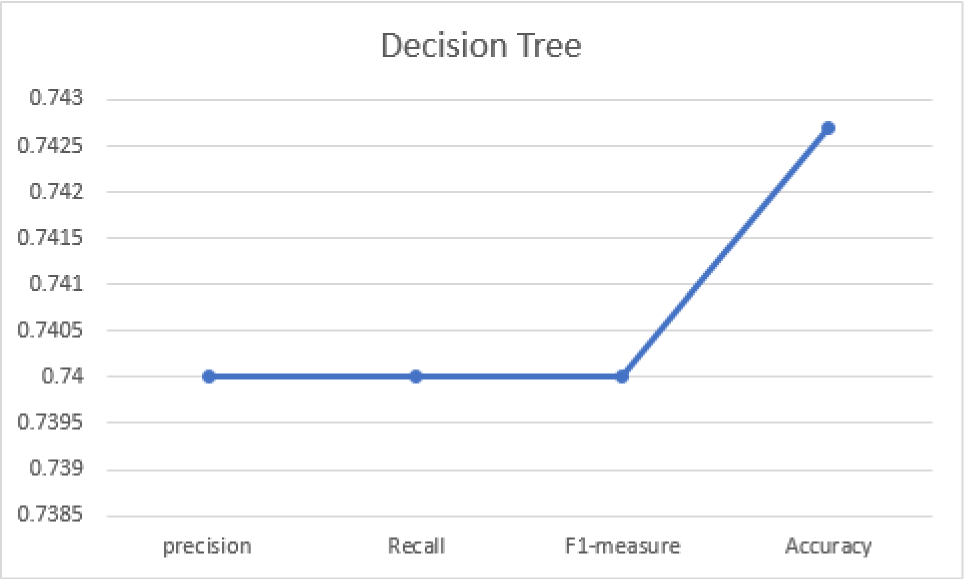


Fig: 14 shows that the actual plot of the Decision Tree algorithm for given table 10 values

Algorithm	Accuracy
Bernoulli	0.786
perceptron	0.84
Logistic Regression	0.85
Decision Tree	0.77
LSTM	0.94

Table:11 shows the Machine learning algorithm with LSTM. LSTM gives better accuracy of the chosen four important algorithms.

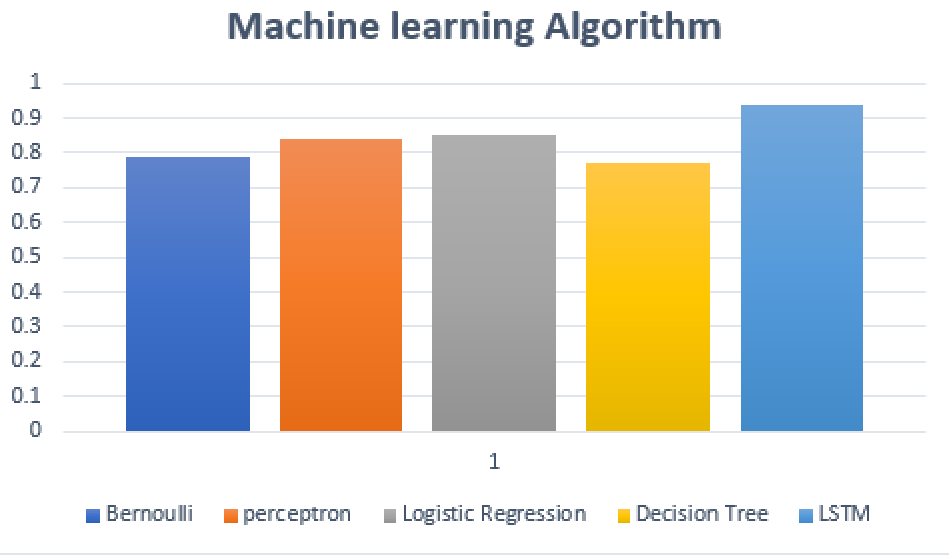


Fig 15: Here is the important plot of the Machine Learning algorithm and its Accuracy by the plot

Chapter 6

FUTURE WORK

In Future work, expectation based strategies will be actualized with existing approach. More highlights will be extricated to handle the certain opinion examination. It can be keep tuning existing neural arrange by choosing diverse parameters such as learning rate, regularization rate, implanting measure etc. to assist move forward expectation exactness. Other than, it can be too attempt a few state-of-art traps for neural arrange preparing such as Bunch Normalization [12], Arbitrary Rummage around for Hyper- Parameter Optimization [13]. Future work moreover incorporates applying the unused system and explanation construction to other dialects and upgrading it with data approximately subjects or occasions, supposition holders, and explanations for etymological wonders like allegory and irony.

Aspect Based sentiment analysis could be a content investigation strategy that breaks down content into perspective term and after that make viewpoint extremity and for diminishing the terms it identifies categories and once more make the category extremity at that point designates in each content which one is estimation level like positive, negative or impartial.

In CNN, it can offer assistance to induce way better include to memorize the errand by substituting convolution and sub-sampling. By utilizing progressed concepts like neural nets, this framework can be moved forward to supply incredible precision. For progressing the execution, extra eatery audits can be included to construct distributional semantic demonstrate. Including preparing information for angle extraction, viewpoint categorization, and assumption classification step can too move forward the execution of the models. Other than that, it ought to progress preprocess step. A sack of positive and negative words like “delicious”, “bad” etc. were utilized to rate the audits based on word score comparison. Audit that has the most noteworthy score was ranked at to begin with position and so on. Hence a requested list is ready and given as yield to the client. In future it'll be comparing positioning strategy with different other accessible methods.

CONCLUSION

In this study, it has been classified an Amazon food review data set in to two categories, positivity and negativity by using some machine learning algorithms. Nowadays, the use of binary classification is increasing. By binary classification detecting positive food review is really useful. The goal of binary classification is information focuses into one of two buckets: 0 or 1, true or untrue, to outlive or not to outlive, blue or no blue eyes, etc. In this research, food review classification done by using different Machine Learning Algorithm with sentiment analysis and feature selection. In this study, it used Kaggle kernel tool which is online based on tool. This study used amazon food review dataset for binary classification. And the dataset consists of total 568454 data. The result based on accuracy for basic machine learning algorithm. With preprocessing, Bernoulli showed 78%, Perceptron showed 84%, Logistic showed 85%, Decision tree showed 77%. And Without preprocessing Bernoulli showed 76%, Perceptron showed 81%, Logistic Regression showed 83%, Decision tree showed 74%. It also showed that LSTM which is 95.85%, gives the better performed than other algorithms. It applied LSTM on single reviews and said some positive case positivity 95.85% and the negative word for negativity 79.08%. All this results were graphically showed for better understanding. Also hope that it will be helpful for people.

Bibliography

[1] S. Garcia and P. Yin, “User Review Sentiment Classification and Aggregation,” no. 4, pp. 0–5, 2015.

[2] Z. Zhou and L. Xu, “Amazon Food Review Classification using Deep Learning and Recommender System,” *Stanford Univ.*, pp. 1–7, 2009.

[3] J. Wu and T. Ji, “2016_Deep Learning for Amazon Food Review Sentiment Analysis,” pp. 1–8, 2016.

[4] C. Wu, A. Ahmed, and A. J. Smola, “Explaining reviews and ratings with PACO : Poisson Additive Co-Clustering,” pp. 1–20, 2015.

[5] J. Mcauley and B. Russell, “From Amateurs to Connoisseurs : Modeling the Evolution of User Expertise through Online Reviews,” 2013.

[6] D. Tsukiyama, “Amazon Fine Food Reviews … wait I don ’ t know what they are reviewing Dataset,” 2012.

[7] “AMAZON FINE FOOD REVIEWS – DESIGN AND IMPLEMENTATION OF AN AUTOMATED CLASSIFICATION SYSTEM,” no. May, 2019.

[8] C. Prabhavathi, “MACHINE LEARNING MODEL FOR CLASSIFYING L _ TEXT USING NLP (AMAZON PRODUCT REVIEWS),” vol. 6, no. 04, pp. 161–178, 2019.

[9] V. Bhati and J. Kher, “Survey for Amazon Fine Food Reviews,” pp. 601–603, 2019.

[10] Y. Chu, “Predicting Rating of Amazon Fine Food from Reviews CSE 190 Assignment 2,” 2012.

[11] C. Zheng, “Rating prediction on Amazon Fine Foods Reviews.”

[12] F. Bornebusch *et al.*, “Aspect-based sentiment analysis,” *Lect. Notes Informatics (LNI), Proc. - Ser. Gesellschaft fur Inform.*, vol. P-232, pp. 2389–2400, 2014.

[13] B. Lu, M. Ott, C. Cardie, and B. K. Tsou, “Multi-aspect sentiment analysis with topic models,” *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 81–88, 2011.

[14] Anshuman, S. Rao, and M. Kakkar, “A rating approach based on sentiment analysis,” *Proc. 7th Int. Conf. Conflu. 2017 Cloud Comput. Data Sci. Eng.*, pp. 557–562, 2017.

[15] D. Ekawati and M. L. Khodra, “Aspect-based sentiment analysis for Indonesian restaurant reviews,” *Proc. - 2017 Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2017*, 2017.

[16] S. Garcia and P. Yin, “User Review Sentiment Classification and Aggregation,” no. 4, pp. 0–5, 2015.

[17] A. You, M. A. Y. Be, and I. In, “Aspect-based sentiment analysis to review products using Naïve Bayes Aspect-based Sentiment Analysis to Review Products Using Naïve Bayes,” vol. 020060, no. August, 2017.

[18] Z. Zhou and L. Xu, “Amazon Food Review Classification using Deep Learning and Recommender System,” *Stanford Univ.*, pp. 1–7, 2009.

[19] J. Wu and T. Ji, “2016_Deep Learning for Amazon Food Review Sentiment Analysis,” pp. 1–8, 2016.