Investigation of Different Machine Learning Approaches for Analyzing Human Attitude

Ifthakhar Ahmed

ID: 2015-2-60-028

Golam Mostafa

ID:2015-2-60-005

A thesis submitted in partial fulfillment of the requirement 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 Dr. Mohammad Rezwanul Huq, Assistant Professor, Department of Computer Science and engineering, East West University. We also declare that no part of this thesis/project has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

……

(Dr. Mohammad Rezwanul Huq)

Supervisor

Signature

……

(Ifthakhar Ahmed)

ID:2015-2-60-028

Signature

……

(Golam Mostafa)

ID:2015-2-60-005

ABSTRACT

Sentiment analysis is widely used in data science, where data is used and analyzed, which is available in various social media and internet. Sentiment analysis is a qualitative processing of text data that extracts and defines subjective details in the source material and allows a company or something like that to recognize the feeling of its brand or products or services when tracking online conversations and feedback. Social media analysis is usually limited to simple analysis of sentiment or count based metrics. Everyone is articulate one way or another in in this time. Most social media and android apps like Facebook or WhatsApp or Twitter have a lot of information which is accessible in this highly developed and re imagined world. Twitter is one of the most common and international networks. Twitter is that kind of social media where many users can express their opinion and feelings through small tweets. These tweets can be analyzed using different machine learning algorithms. Twitter sentiment analysis is a very popular research work now. Most of the work is on two types of sentiment detection, it can be either positive or negative. This paper includes neutral sentiment. So, the proposed idea is to find out a tweet is positive or negative or neutral with a better accuracy. In this paper tweeter data has been encoded using label encoder OneHot encoder. After that applying different types of pre-processors, we have feed that numeric form to the machine learning classifier algorithms. Although our accuracy is quite low, in future we shall develop the algorithm for better accuracy.

Acknowledgements

Firstly, we want to express our profound gratitude to the all-powerful God because of His blessings upon all of us.

Next, we are thankful to our supervisor "Dr. Mohammad Rezwanul Huq, who gave us the opportunity, motivation and introduced us to the "Data Science" sector and without whom this research was not possible. His motivation, visionaries, insightful tips, advice and unforgettable help at all stages of our BSc research were acknowledging and important. His desire to correctly answer our every question is something that we have found very important, and we'd try to emulate if we do get the chance.

There are also many other individuals who showed us their continuous support and encouragement knowingly or unknowingly linked to our educational life in different ways. We will remember them in our heart and hope to find a more appropriate place to acknowledge them in future.

In the end, we would like to thank our parents, our friends and our brothers and sisters for their support.

Ifthakhar Ahmed

December 2019

Golam Mostafa

December 2019

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF Algorithms

CHAPTER 1

INTRODUCTION

Technology is advancing day by day and sentiment analysis in one of them. It is vastly used nowadays. Sentiment analysis is a way to identify people's opinion available in social media and internet.

1.1 Introduction

Sentiment analysis is a very important technology. For instance, if anyone want to open an online business, people's review is very important. It will help him to understand how his products are or what a buyer wants. There will be lots of products and lots of reviews. So, it will be not possible for him to read one by one review and understand what it says. Now sentiment analysis will help to give him an overall idea. This is one of the reasons why sentiment analysis is important.

Sentiment analysis not only helps a company to understand how they're doing with their customers; it also gives them a better idea of their position against their competitors.

Like online business we can also identify people's opinion from social networking sites like Facebook, Twitter using sentiment analysis. The use of microblogging is increasing every day. Many small messages contained by twitter is created by the internet user. Twitter is a fast-growing social media where people can express their opinion using small tweets. Twitter sentiment analysis is a very popular research work now.

A paper shows that how to use Twitter as a corpus for sentiment analysis and opinion mining. The corpus they used can be arbitrary large. The dataset contains 300000 corpus of text posts which split automatically between three sets of texts (positive, negative and neutral). They perform a linguistic analysis of their corpus to show how to build a sentiment classifier that uses the dataset.

In this paper Twitter data was analyzed using different machine learning algorithm to identify people's opinion. People can have three types of opinion. They are positive, negative and neutral. Here a tweet will be an input. The input size is not pre-defined here. The tweet can be if possible. After having input the system will analyze the sentence and predict the attitude of the tweet. Here the tweet was classified in three different class like positive, negative and neutral as we mentioned earlier. Some example of input and output is described below.

Example of Positive Tweet:

1. @VirginAmerica done! Thank you for the quick response, apparently faster than sitting on hold ;)
2. @united I appreciate your efforts getting me home!

Example of Negative Tweet:

1. @united is the worst. Nonrefundable first-class ticket? Oh, because when you select Global/FC their system auto selects economy w/upgrade.
2. @united I will not be flying you again.

Example of Neutral Tweet:

1. @VirginAmerica my driver's license is expired by a little over month. Can I fly Friday morning using my expired license?
2. @VirginAmerica any plans to start flying direct from DAL to LAS?

1.2 Motivation

The inspiration came from a sentiment analysis system based on twitter, where they were developing a template to divide the tweet into positive or negative and neutral feelings. For two classification tasks, they create models: a binary one to define positive and negative feelings, and a three-way one to classify

the dataset into positive, negative and neutral feelings. They use three forms of models, perhaps even a model based on unigram, a model based on functionality and a model based on the tree kernel. Their feature-based model, which uses just 100 features, maintains similar accuracy as the unigram model, which uses more than 10,000 elements. This paper also shows that now the tree kernel model performs approximately and the best system-based models, although thorough software engineering is not needed. For their experiments, they use personally annotated Twitter data. A variety of the functions are based on preceding word polarity. An inspiration came from work by Agarwal et al. (2009) to obtain the prior polarity of terms. Using the Language Dictionary of Affect and use WordNet to extend it. It comprises of 8000 English words that assign a pleasurable score between 1 (Negative)-3 (Positive) to each word. Words with polarity below 0.5 are considered negative, greater than 0.8aspositiveandtherestasneutral. To combine many types of features in one succinct versatile representation, we format a tree recognition of tweets. Using a Partial Tree kernel proposed by Moschitt for the very first time to calculate the correlation between two trees. The PT tree kernel generates and compares all possible sub-trees. Such subtrees provide subtrees where, although order is conserved, non-adjacent fences adjacent by removing other branches. There are 50 types of characteristics. For the entire tweet and last one-third of the post, calculating these features. We have 100 features in total. Every feature of Polar and Non-polar will be further divided into two categories: POS and the Other. POS refers to apps collecting word parts statistics and many other relates to all other interface types. Tree kernel-based model reaches the highest precision of 60.60 percent accuracy compared to the others. [5]

They chose English for the work intent but also because it requires Twitter API protocol, their approach can be adapted to other language. Initially, the word frequency distribution in the corpus was tested. They use the law of Zapf to distribute the quantities of the word. Second, they marked all the articles in the corpus using Tree Tagger approach for English. The obtained data set is being used to derive features to train our classifier of feelings. We used the existence of a n-gram as a binary function while retrieving basic information. To search the training dataset, they removed URL references and usernames from

Twitter. By breaking it into spaces and punctuation, they segment text and form a bag of letters. The tokenization is done in this section. We removed items from the bag of words as stopwords. Then, to classify the dataset, an N-grams is built. They discard N-grams to enhance the accuracy. As a result, to achieve greater accuracy, bi-grams are better conducted. [4]

1.3 Objective

The main research targets are as follows:

- Collecting the real time tweeter data using API.
- Applying different classification algorithms on the dataset to find the accuracy and efficiency.
- Comparing the accuracy among the classifiers

1.4 Contribution

1. It is a tokenization technique which have stopwords implementation. It can automatically analyze the data and predict the sentiment.
2. This will define human's sentiments with the help of machine. Machine learning methods have a major impact on the day-to-day interaction of individuals in data science.
3. Most of the paper have detected two sentiments. It is either positive or negative. In this paper three types of sentiment were detected. Positive negative and neutral.
4. To detect the sentiment, some well-known classification approaches were applied. For example, the output of this proposed method was compared with Neural Network Classifier, k-Neighbors Classifiers, Logistic Regression, Nearest Centroid, Naive Bayes.

1.5 Outline

Chapter 1: Introduces sentimental analysis and the motivation as well as objective.

Chapter 2: This chapter illustrates the background of our proposed methods and the related works.

Chapter 3: Chapter 3 shows the architectural view of our proposed method.

Chapter 4: This chapter shows the statistical analysis of our whole dataset.

Chapter 5: Chapter 5 describes the implementation process, algorithms that are used and pseudocodes.

Chapter 6: This chapter analyzes the results obtained from our proposed methods.

Chapter 7: The final chapter summarizes the overall work that we have done and explains the future works that we need to focus on.

CHAPTER 2

LITERATURE REVIEW

2.1 Background

2.1.1 API:

API is a software application deployment tool for application system interface. Mainly it provides a way to interact with software components. Through supplying all the structures, a strong API makes it much easier to develop a system. A developer then brings together the blocks. [1]

It connects to the Internet using an application and sends information to a database. The server then retrieves the information, interprets it, performs the actions required and sends it back to the phone. Then the program interprets the data and displays the information in a readable manner. These are all happening with the aid of the API.

2.1.2 Tweepy:

Tweepy allows you bypass a lot of those low-level details. It is an open source package. Twitter's developer website has great documentation to see the type of data you can access. Tweepy's documentation has code and some basic documentation for the Tweepy module. [2]

2.1.3 Stopwords:

One of the major forms of pre-processing is to filter out useless data. In data science useless words are referred to as stop words.

The messy existence (abbreviations, unusual types) of information usually affects every dataset. A common technique for reducing textual data noise is to delete stopwords using pre-compiled sets of stopwords. The effectiveness of eliminating stopwords has been discussed in the form of sentiment analysis over the past few years. [3]

2.1.4 Neural Network Classifier:

Neural network has units, or it can be said that layered neurons that can convert an input vector into some output. Each unit takes an input and relates a function to it, moving the output to next layer. Such networks are primarily known as feed-forward. Neural network has units, or it can be said that layered neurons that can convert an input vector into some output. Each unit takes an input and relates a function to it, moving the output to next layer. Such networks are primarily known as feed-forward. [4]

Author in [14] has used the neural network algorithm as a classifier. They have predicted the Bankruptcy from the bank data. Their accuracy was quite impressive, in this case they used three set of datasets to apply the algorithm. On the other hand, author in [15] also use the neural network as a classifier. Their main purpose was to classify the ECG signal effectively.

Neural networks have various problems with an application. This range of feature is a pattern recognition, we're going to consider this here.

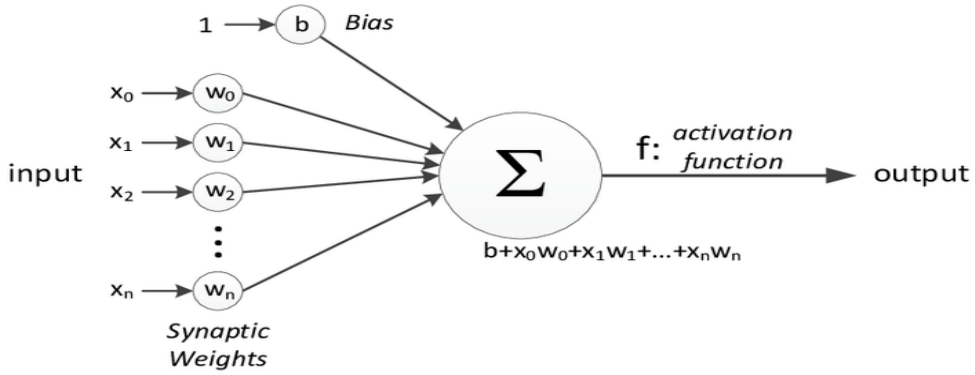


Figure 2.1.1: Neural Network

2.1.5 K- Nearest Classifier:

k-nearest classifier algorithm is a simplest and effective rule for pattern classification. It is assigned to a class level to each query pattern which is compared with the most near centroid value of the classifier. The algorithm works well in the small size cases compared with the large size data including state-of-the-art KNN algorithm. [5]

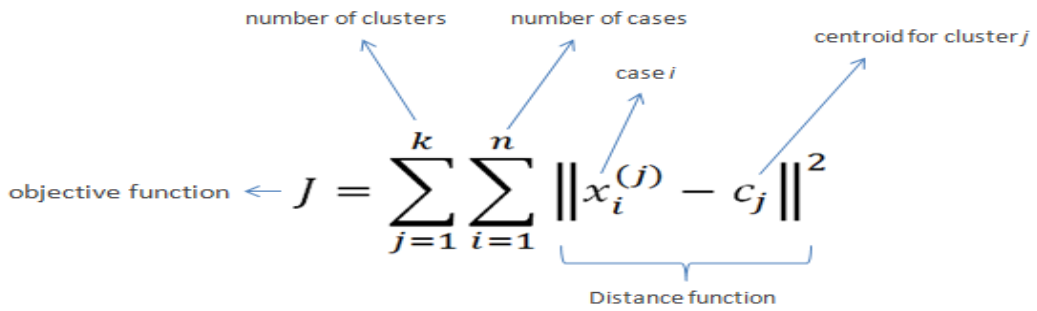


Figure 2.1.2: Function of K-nearest Classifier

2.1.6 Logistic Regression:

In case the dependent factor is dichotomous (binary), logistic regression is the correct regression analysis to perform. It is a statistical method, like all regression analyses. It is used to characterize data and illustrate the relationship with one dependent binary variable and one or more independent nominal, ordinal, interval or ratio-level variables. [6]

$$\text{logit}(p) = \ln\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 x_{i2} + \beta_2 \cdot x_{i2} + \_ + \beta_p \cdot x_{in}$$

for i = 1...n .

Figure 2.1.3: Equation of logistic Regression

2.1.7 K-Nearest Centroid:

It is a classification that assigns objects whose mean or centroid is nearest to the observation to observations of the tag band. The centroid presents every class, with test samples assigned to the nearest centroid level.

A new proposed k-Nearest Centroid classification rule with two heuristic changes was introduced by the author in [18]. Such options used in the training set for the geometric and distance distribution of models to estimate the category tag of a given sample. Author in [19] suggested a local mean-based

neighborhood classifier k-nearest centroid which allocated a class tag with the nearest centroid mean vector to each query sequence to improve classification efficiency. In addition to considering the proximity and geographic distribution of k neighborhoods, the proposed scheme also uses the local mean vector of k neighbors from each group to make classification decisions.

Nearest-centroid classifiers were widely used in various high-dimensional applications, in recent genomics. The researcher in [7] has illustrated on their patient list and has been able to acquire a sufficient accuracy.

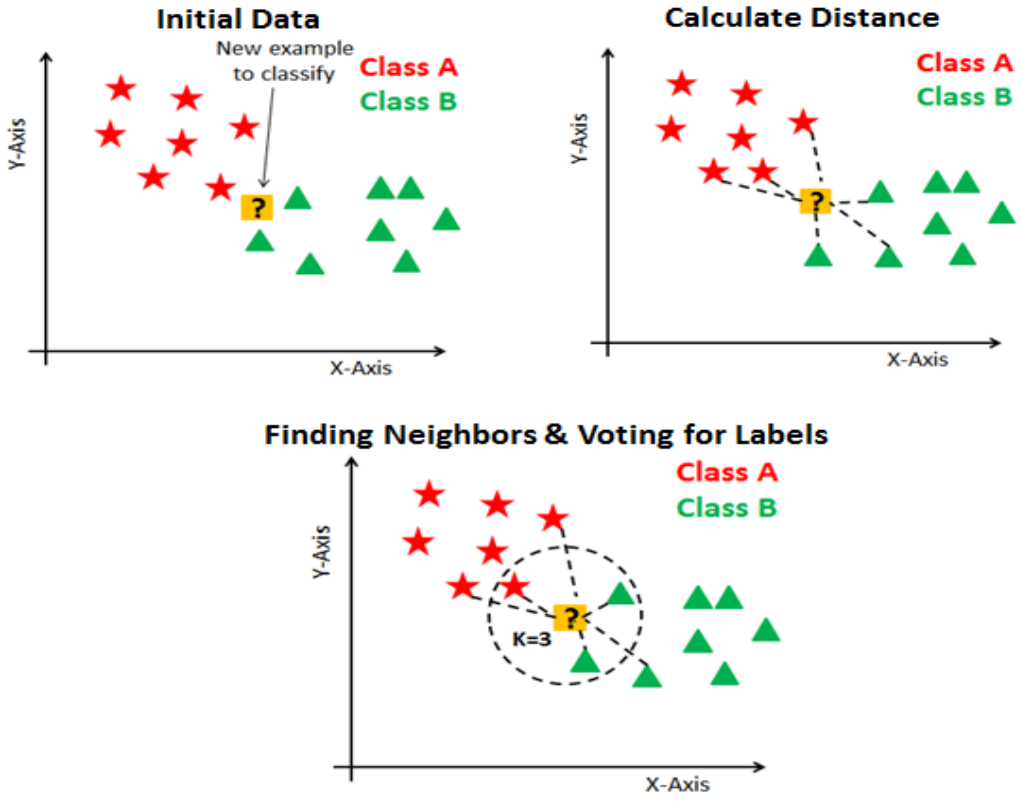


Figure 2.1.4: K-Nearest Centroid

2.1.8 Naive Bayes:

Naive bayes algorithm is based on theory of bayes and consisting a set of classification algorithms. It contains of many algorithms where they all share a common law. Each pair of characteristics is distinct from each other. Naive bayes algorithms work well for text classification. [8]

The researcher in [20] demonstrated a simple approach to fake news detection that used a nNaive Bayes classifier which was used as a software and checked against a set of data of Facebook news posts. On the test set, they achieved 74 percent recognition accuracy. The writer in [21] suggested three equivalents from Bayes in which it turned out that the classical NB algorithm with the Bernoulli event model is identical to the Bayesian counterpart.

2.1.9 Support vector machine:

Support Vector Machine is a classifier that is described by a separate hyperplane. It's a supervised type of learning. In two-dimensional space, a hyperplane is a line splitter in two sections where there are two sides in each class [9]. Author in [16] represents a survey of the monitoring of machine condition that uses support vector machine to diagnose errors. It seeks to summarize recent SVM research into the monitoring and diagnosis of machine condition. The author is proposed in [17] used Support Vector Machines as a new method of machine learning to predict the α-turn types in proteins.

$$h(\boldsymbol{x}_i) = sign(\sum_{j=1}^{s} \alpha_j \; y_j \, K(\boldsymbol{x}_j, \boldsymbol{x}_i) + b \;)$$

$$K(\boldsymbol{v}, \boldsymbol{v}') = \exp(\frac{||\boldsymbol{v} - \boldsymbol{v}'||^2}{2\gamma^2})$$

Figure 2.1.5: SVM equation

2.2 Related Work

There is a lot of work on understanding the emotions. A paper attempted to improve performance with the accuracy of the methodology of sentiment analysis. This shows that the proposed team classifier works much better than stand-alone classifiers with majority participating ensemble classifier. As part of their work, the role of pre-processing data and representation of features in sentiment analysis techniques is also shown. [13]

Extracting and selecting text features is the first step of the Sentiment analysis problem. Some different approaches can be used, such as extracting information including Terms Presence and Frequency in text details, Speech Parts (POS), Words and Phrases of Opinion, Negation, etc. Point wise mutual information is also valuable for the analysis of the data set, since it is a formal way of modeling the mutual information between features and classes. There are two approach to analyze the sentiment and give a feedback. The approaches are Lexicon based approach and Hybrid approach. The lexicon-based approach depends on findings the opinion lexicon which is analyze the text. [10]

For short extracting the features from a short text, a system can be used in convolutional neural network. Based on the short text, by activation values of an inner layer of deep convolutional neural network it can be solved. [11] After training a deep convolutional neural network the data can be classified, instead of using convolutional neural network as a classifier some values from the hidden layer can be used as features for a much more advanced classifier, and this way gives a superior accuracy. [11] Using CNN as only a classifier gives lower accuracy than the accuracy has been found from CNN which can be used to extract trainable features for the SVM classifier. Two different method named feature level fusion and decision level fusion can also be used for better result. [11]

The dataset can be tuning by a subset which is known as labels. And the subset can be found in an unsupervised learning, where each layer can automatically learn features. [12] A single kernel vector classifier can adapt different modalities and it can give a higher accuracy rate. To train the machine for multi-layer system the gradient of the total energy function with respects to the weights of all layer have to be compute. [12] The approximate higher likelihood contrastive divergence mechanism can be useful. Logistic regression model is work better for normal distribution to learn the binary hidden neurons which is visible. The Recurrent Neural Network, Convolutional Neural Network and Multi Kernel Learning can be integrated to create a new model named CRMKL. Which performs better for feature analysis. [12]

CHAPTER 3

3.1 PROPOSED METHOD

Since, the number of human being are increasing day by day. Scientists are determined to analyses the sentiment from the various kinds of people using different kinds of machine learning algorithm. Some algorithms are costly and some are efficient in order to determine the sentiment of the sentence, or of a human. The increasing number of human is a fact to analyze the sentiments. So, people may think why we need to analyze the sentiment. By analyzing ones' sentiment it can be determined what the motivation of that person is either he can occur a crime or not. Or the people from various political party can analyze about the people either they are on their side or not. However, proposed of this work consists of the following segments.

- Extracting data from dataset file and splitting them into train (80%), test (20%) .
- Applying LabelEncoder [23] and OneHotEncoder [24] on the training, testing, to convert them into corresponding binary value to feed into machine learning approaches as input.
- Preprocessing data using tokenization, stop word filtering and lemmatizing the word.
- Applying classification approaches on the converted data.
- Further applying classification approaches Support Vector Machine, Logistic Regression, Nearest Centroid, Logistic RegressionCV, Naïve Bayes for performance comparison purpose.
- Finally, performances of all the classification approaches have been analyzed, compared and demonstrated based on accuracy, specificity, memory usage.

However, complete infrastructure of our proposed approach has been represented graphically in Figure 3.

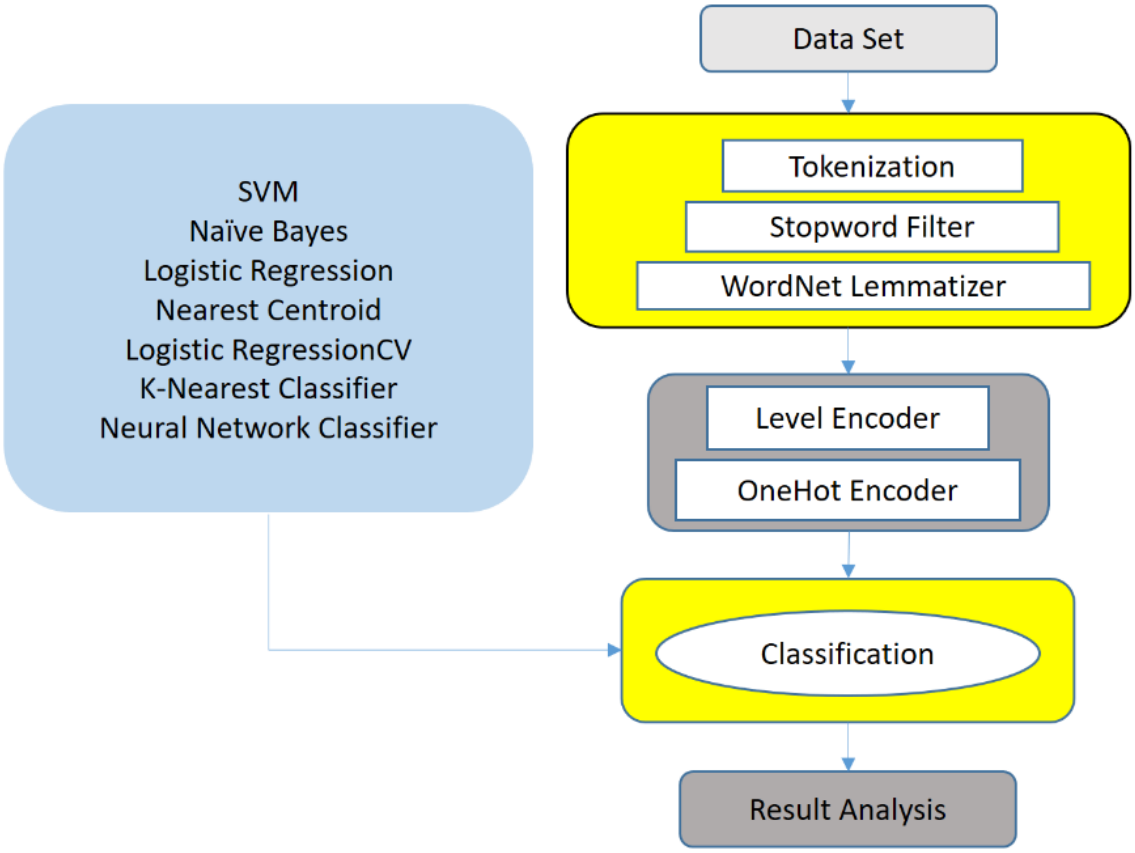


Figure 3.1: Infrastructure of the proposed method

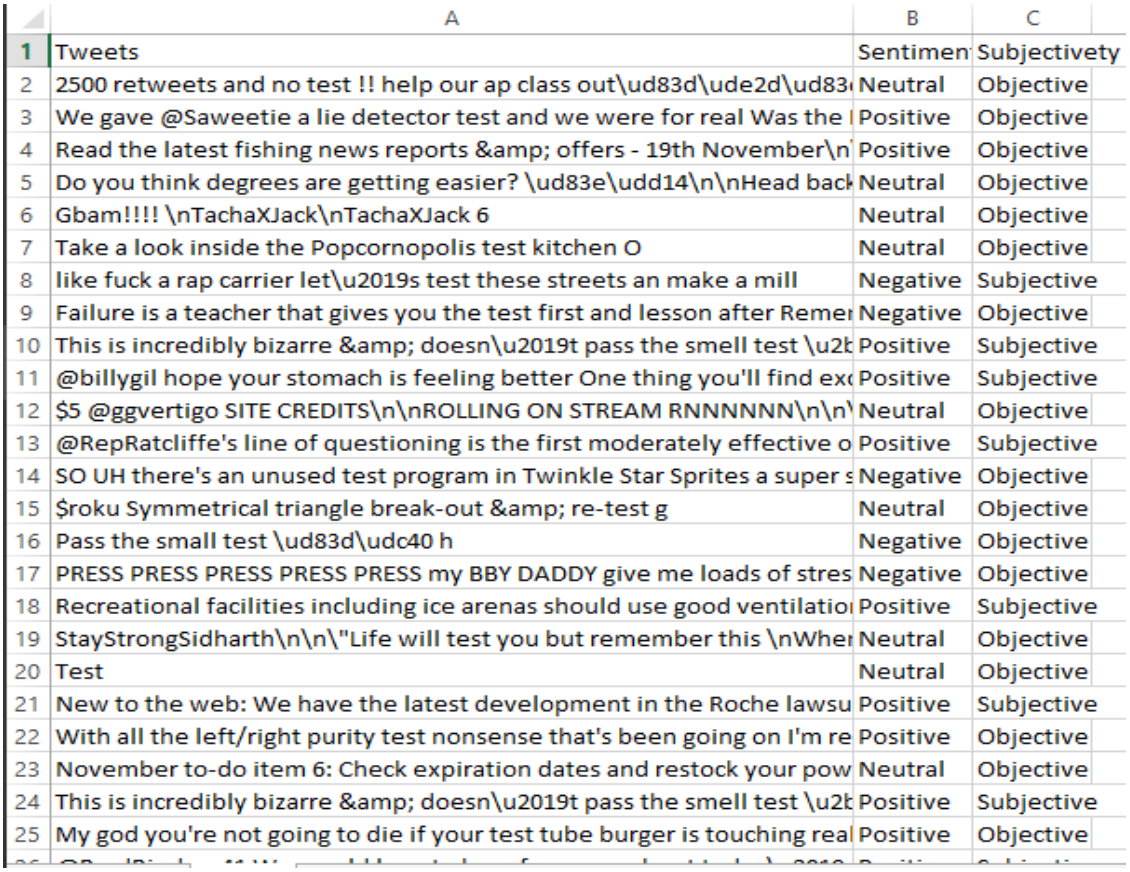
CHAPTER 4

DATASET OVERVIEW

4.1 Attributes Description

The dataset for "Sentimental Analysis" has been obtained from Tweeter developer account. In the dataset there are three columns and 0.5 million rows, means there are 0.5 million of data of the tweeter users. From the dataset the retrieval data or mining data would be sentiment and subjectivity. In the very beginning the sentiment was given as a numeric form such as (-1 to +1). It means if the polarity becomes (-ve) range it means that the sentence is negative or negative sentiment. On the other hand if the polarity become (+ve) then it means it positive sentiment. If it becomes zero, then the sentence stands for the neutral. Another column stand for the subjectivity so, it works such way as the sentiments. It also has the range of it. When it becomes positive it will be subjective occurrence and of the polarity becomes negative it means the objective. In the subjectivity column the subjective means the actual fact occur and the object means the opinions of others. That's how the dataset has been distributed.

- POSITIVE – the sentence has positive sentiment
- NEGATIVE – the sentence has negative sentiment
- SUBJECTIVETY – the fact that actually occurs
- OBJECTIVETY – rather than just an opinion



	A	B	C
1	Tweets	Sentimen	Subjectivety
2	2500 retweets and no test !! help our ap class out\ud83d\ude2d\ud83	Neutral	Objective
3	We gave @Saweetie a lie detector test and we were for real Was the	Positive	Objective
4	Read the latest fishing news reports & offers - 19th November\n	Positive	Objective
5	Do you think degrees are getting easier? \ud83e\udd14\n\nHead back	Neutral	Objective
6	Gbam!!!! \nTachaXJack\nTachaXJack 6	Neutral	Objective
7	Take a look inside the Popcornopolis test kitchen O	Neutral	Objective
8	like fuck a rap carrier let\u2019s test these streets an make a mill	Negative	Subjective
9	Failure is a teacher that gives you the test first and lesson after Remer	Negative	Objective
10	This is incredibly bizarre & doesn\u2019t pass the smell test \u2b	Positive	Subjective
11	@billygil hope your stomach is feeling better One thing you'll find exc	Positive	Subjective
12	$5 @ggvertigo SITE CREDITS\n\nROLLING ON STREAM RNNNNNN\n\n'	Neutral	Objective
13	@RepRatcliffe's line of questioning is the first moderately effective o	Positive	Subjective
14	SO UH there's an unused test program in Twinkle Star Sprites a super s	Negative	Objective
15	$roku Symmetrical triangle break-out & re-test g	Neutral	Objective
16	Pass the small test \ud83d\udc40 h	Negative	Objective
17	PRESS PRESS PRESS PRESS PRESS my BBY DADDY give me loads of stres	Negative	Objective
18	Recreational facilities including ice arenas should use good ventilatio	Positive	Subjective
19	StayStrongSidharth\n\n\"Life will test you but remember this \nWher	Neutral	Objective
20	Test	Neutral	Objective
21	New to the web: We have the latest development in the Roche lawsu	Positive	Subjective
22	With all the left/right purity test nonsense that's been going on I'm re	Positive	Objective
23	November to-do item 6: Check expiration dates and restock your pow	Neutral	Objective
24	This is incredibly bizarre & doesn\u2019t pass the smell test \u2b	Positive	Subjective
25	My god you're not going to die if your test tube burger is touching real	Positive	Objective

Figure 4.1: Total Attributes in dataset

4.2 Occurrence of each significance

Among 505932 significances in the dataset, 49% of the whole dataset is classified as positive sentiment. After that, 29% is classified as Neutral. Moreover, 22% are classified as Negative. Therefore, for 49% people are bringing their sentence as positive however the number is less of the negative statements is 22%. Each significance and their appearances in the dataset have been illustrated in figure 4.2.

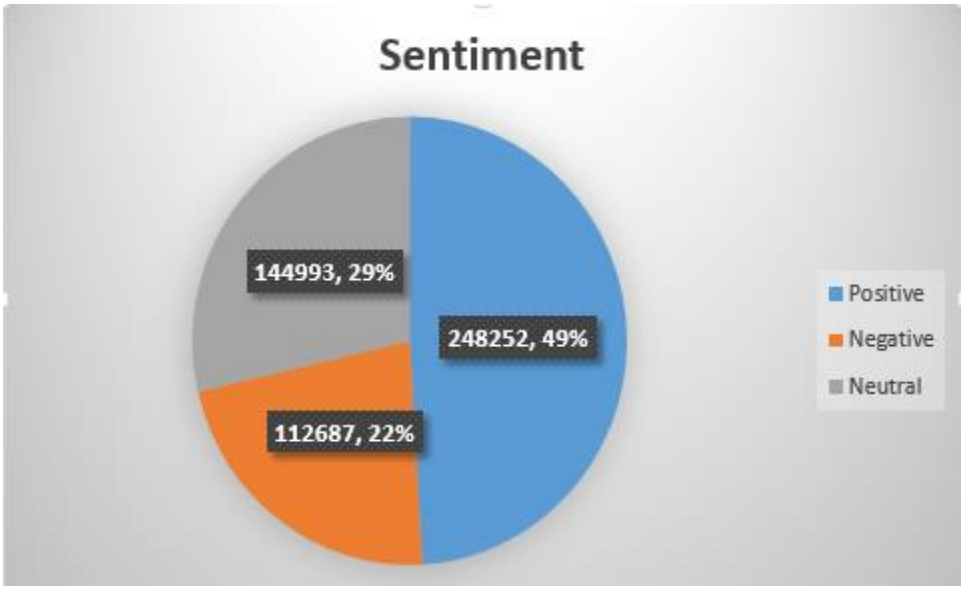


Figure 4.2: Occurrence of every significance in dataset

CHAPTER 5

IMPLEMENTATION

Actually, only one approach has been focused and implemented in this work. Only is classification approaches. In case of classification approach, we have mostly concentrated on different types of classification algorithm such as Neural Network, Logistic RegressionCV, Naïve Bayes, Nearest Centroid, Logistic Regression etc.

5.1 Classification Approaches

All the parameters with specific values are referred in table 5.1. Those parameters can be assigned to regenerate the findings we have acquired here in this work.

Table 5.1: **Required parameters and values**

Parameter	Value	Parameter	Value	Parameter	Value
Iterations	3000	No of Layers	5	No of Weights	10
Train_Size	85%	Nodes Per Layer	10	No of Biases	10
Learning Rate	0.01	Activation Function	Tanh	Initial Value (Biases)	1
Loss Function	BinaryCrossEntropy	Output Activation Function	Sigmoid		
Range of Weights	(-1, 1)	Initial Value (Weights)	1		

Additionally, for comparative study of the performance with different types of algorithms, we have implemented Support Vector Machine, Logistic Regression, Naïve Bayes, Nearest Centroid, Logistic RegressionCV, Neural Network and K-Nearest Neighbor.

Algorithm 5.1: NN (Pseudocode):

Preprocessing:

1. LabelEncoder
2. oneHotEncoding

Steps:

begin

Training stage:

h<-0

step 0: initialize

step 2: quantize

step 3: create random table and weight tables

while h< $epoch_{max}$ do

for j = 1,2,…..,D_{train} do

step 3: generate address table

step 4: calculate actual output

step 5: learning algorithm

endfor

step4:evaluate classification error h<-h+1

endWhile

Testing stage:

Step7: measure classification error

end

End loop.

Algorithm 5.2: Logistic Regression (Pseudocode):

– **Input:** Features $\mathbf{X}$, Target $\mathbf{Y}$
– **Output:** Parameter $\theta = [\theta_0, \theta_1, \cdots, \theta_d]$

$\theta_0, \theta_1, \cdots, \theta_d \leftarrow$ HE.EncryptNumber(0)
$iter \leftarrow 25$
$\alpha \leftarrow$ Decide learning rate
$b \leftarrow$ Set b to an integer where $\frac{\alpha}{n} \approx 2^b$
repeat ▷ Compute $\sum_{i=1}^{n} (h_\theta(\mathbf{X}^{(i)}) - \mathbf{Y}^{(i)})\mathbf{X}_j^{(i)}$
  for $j = 0$ to d **do**
    $\hat{grad}_j \leftarrow$ HE.EncryptNumber(0)
    for $i = 1$ to n **do**
      $\hat{s}_j \leftarrow$ HE.Sigmoid$(\theta^T \mathbf{X}^{(i)})$
      $\hat{s}_j \leftarrow$ HE.Subtract$(\hat{s}_j, \mathbf{Y}^{(i)})$
      $\hat{s}_j \leftarrow$ HE.Multiply$(\hat{s}_j, \mathbf{X}_j^{(i)})$
      $\hat{grad}_j \leftarrow$ HE.Add$(\hat{grad}_j, s_j)$
    end for
    $\hat{grad}_j \leftarrow$ HE.RightShift$(\hat{grad}_j, b)$
    $\theta_j \leftarrow$ HE.Subtract$(\theta_j, \hat{grad}_j)$
  end for
until $iter$ times
return θ

Figure 5.1: pseudocode of Logistic Regression

Algorithm 5.3: Pseudocode of SVM

Data: Dataset with p* variables and binary outcomes.

Output: Ranked list of variables according to their relevance.

Find the optimal values for the tuning parameters of the SVM models:

Train the SVM models;

p<-p*;

while p >=2 do

SVMp <- svm with the optimized tuning parameters for the p variables and observations in Data;

wp <- calculate weight vector of the SVMp(wp1,.....,wpp);

rank.criteria <- (wp1^2,.....,wpp^2);

min.rank.criteria <- variable with lowest value in rank.criteria vector;

Remove min.rank,criteria from Data;

Rankp <- min.rank.criteria;

p <- p-1;

end

Rank1 <- variable in Data !∈ (Rank2,.....Rankp*);

return (Rank1,....,Rankp*)

Algorithm 5.4: Pseudocode of K-Nearest Centroid

k-Nearest Centroid

Classify(X,Y,x) // X:training data, Y:class labels of X, x:unknown sample

for i=1 to m do

Compute distance d(Xi,x)

end for

Compute set I containing indices for the k smallest distances d(Xi,x)

return majority label for {Yi, where i?I}

Algorithm 5.5: Pseudocode of Naïve Bayes

```
TrainNB(C,D)
V<-EXTRACTVOCABULARY(D)
N<-COUNTDOCS(D)
for each c?C
do Nc <- COUNTDOCSINCLASS(D,c)
    prior[c] <- Nc/N
    for each t?V
    do Nct <- COUNTDOCSINCLASSCONTAININGTERM(D,c,t)
    condprob[t][c] <- (Nct+1)/(Nc+2)
return V,prior,condprob

APPLYNB(C,V,prior,condprob,d)
Vd <- EXTRACTTERMSFROMDOC(V,d)
for each c?C
do score[c] <- log prior[c]
      for each t?V
      do it t?Vd
          then score[c] += log condprob[t][c]
          else score[c] += log(1-condprob[t][c])
return argmax c?C score[c]
```

CHAPTER 6

RESULT ANALYSIS

This section result and performance of all the considered approaches have been evaluated and analyzed using various performance measurements including timing, accuracy, and memory requirement etc. Moreover, cost per iteration graph have also been described graphically and theoretically for comparative study. Additionally. On the other hand training time and the memory use of time has also been conducted. All these performance evaluation metrics have been concentrated in order to find the superior approach among all the approaches applied here in this work. All the approaches of Comparative studies are as follows:

6.1 Performance Analysis of Classification Approaches

Here, from the table given bellow it can be decided the accuracy of each algorithms. As a result it can comparable to each other to determine which algorithm is most efficient to determine the most accuracies.

In table 6.1 the accuracies of the algorithms are given bellow.

Table 6.1: **Accuracies of the algorithms**

Name of the Algorithm	Depicted Accuracy (%)
Support vector machine	70.00
Naive Bayes	68.16
K nearest Centroid	62.53
Logistic Regression	68.66
Logistic Regression CV	70.33
Neural Network	70.33
KNeighborsClassifier	61.83

Based on the information provided in table, we can generate the graph from the table. That performance of almost all the approaches are quite same rather than a little bit difference. It can be easily notified that Logistic RegressionCV and Neural Network has the 2-3% higher accuracy compared to the other classification alternatives compared in this work. Accuracy comparison has been graphically represented in Figure 6.1.

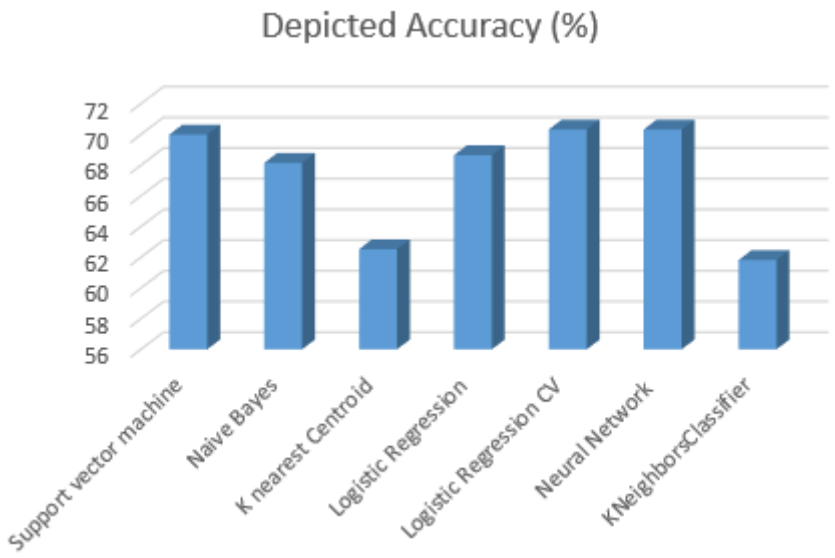


Figure 6.1: Comparing the Algorithms Accuracy

Meanwhile, for implementation of cost function we have considered Binary Cross Entropy mechanism for cost estimation. For better clarification we have extracted cost per iteration graph which is referred in Figure 6.2.

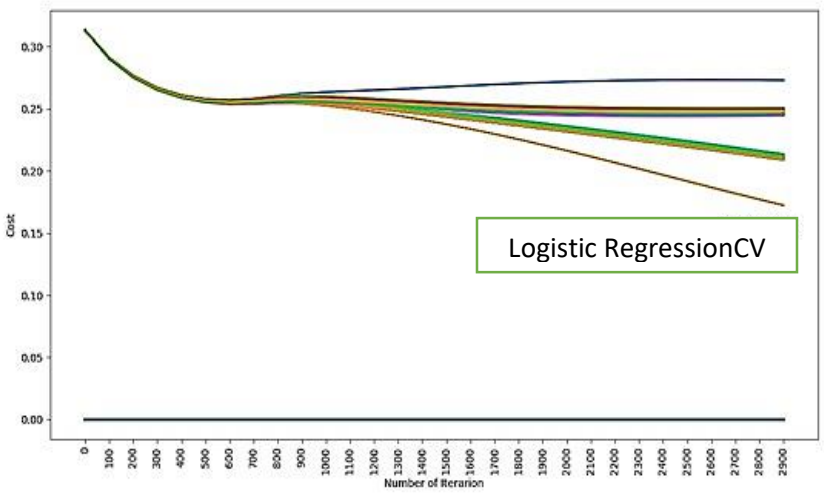


Figure 6.2: Cost per iteration graph

Figure 6.2 demonstrates that Logistic RegressionCV diminishes the cost more than any other approaches in this work which ensures the superior behavior of our approaches. Moreover, for more vivid perception regarding effectiveness and efficiency of our proposed approach, we have manifold as well as observed memory usage and timing requirement in the training phase.

Table 6.2: Comparison of performance of the proposed approach

Name of the Approach	Memory Usage (MB)	Time Required For Training (sec)	Accuracy (%)
Neural Network	413.31	1916.66	70.33
Support vector machine	814.34	20833.33	70.00
Logistic RegressionCV	1001.21	13333.33	70.33
Naïve Bayes	812.34	411.11	68.16
K-nearest Centroid	856.32	399.11	62.53
K Neighbor Classification	756.32	2833.33	61.83
Logistic Regression	701.00	535.833	68.66

Moreover, Table 6.2 refers to total time and memory required while training models on our dataset. In case of time and accuracy our classification approaches it seems to be more promising as well as effective. However, memory requirement of our classification approach is quite higher Support vector machine (SVM) but overall demonstration states that our classification approach reveals the best performance. Graphical representation of Table 6.2 also depicts the same in Figure 6.3.

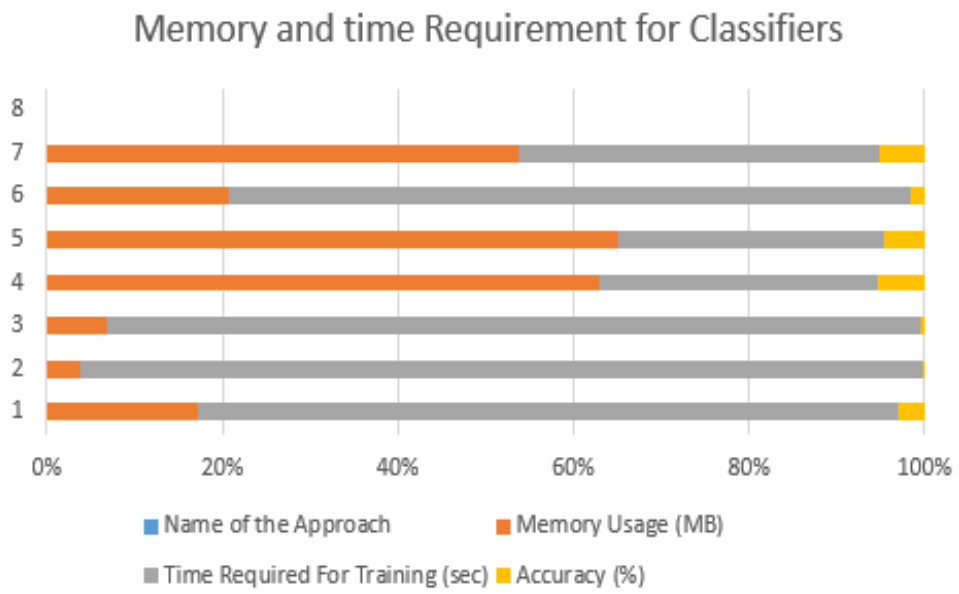


Figure 6.3: Comparison of accuracy, Memory and Time

The experiment reveals that the Neural Network has given most promising result among all. Meanwhile, it took less time and higher accuracy.

CHAPTER 7

CONCLUSION & FUTURE WORK

7.1 Conclusion

Twitter sentiment analysis is a very popular research work now. In this paper, different machine learning algorithms were used to analyze twitter data. Most of the work is on detecting three types of sentiments. They are on positive negative and neutral sentiment. Neutral sentiment is as well as important. In this work accuracy was not quite good. Because the classifiers need to be modified. As we apply the direct algorithm on the dataset that's why accuracy was not quite satisfied. But still there is a very good opportunity to improve this accuracy. So still there is a working scope. This work can be done on even bigger dataset.

7.2 Future Work

In near future we will improve the accuracy and we will work with a large dataset than the present dataset. Our work will be consisting of the both sentiment and subjectivity. As the machine can determine a person's sentiment on the objective or subjective. And we also have two different ideas to work with sentiment analysis. We will make a module using embedded system, where a person will input his voice in different language and after analyzing the sentence the module will reply automatically whether the person's sentiment is. As well we have another idea to work with English language, for instance we wish to work with English newspaper headlines. In that work, the system will automatically detect the sarcasm of English newspaper headlines using different approach.

BIBLIOGRAPHY

[1] Vangie Beal 2019, accessed 30 October 2019,http://www.webopedia.com .

[2] Paulynn Yu Nov 5 2019, accede 30 november 2019, http://towardsdatascience.com.

[3] Oro.open.ac.uk. (2019). On stopwords, filtering and data sparsity for sentiment analysis of Twitter - Open Research Online. [online] Available at: http://oro.open.ac.uk/id/eprint/40666.

[4] Jp Thompson 12 january 2018, accessed 30 November 2019, <https://robots.ox.ac.uk>

[5] Anon, June 2019. Accessed 30 November 2019,https://link.springer.com/chapter.

[6] Jason Brownlee April 1 2016, accessed 30 November 2019, <https://machinelearningmastery.com>.

[7] Bacil, E. D., Mazzardo Júnior, O., Rech, C. R., Legnani, R. F., & de Campos, W. (2015). Atividade física e maturação biológica: uma revisão sistemática [Physical activity and biological maturation: a systematic review]. *Revista paulista de pediatria : orgao oficial da Sociedade de Pediatria de Sao Paulo*, *33*(1), 114–121. doi:10.1016/j.rpped.2014.11.003.

[8] Rushikesh Pupale Jun 16, 2018. Accessed on 30 November 2019, <https://towardsdatascience.com>.

[9] S. Poria, H. Peng, A. Hussain, N. Howard and E. Cambria, "Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis", Neurocomputing, vol. 261, pp. 217-230, 2017. Available: 10.1016/j.neucom.2016.09.117.

[10] S. Poria, E. Cambria and A. Gelbukh, "Deep Convolutional Neural Network Textual Features and Multiple Kernrel Learning for Utterence-Level Multimodal Sentiment Analysis", Conference on Empirical Methods in Natural Language Processing, pp. 2539-2544, 2015whats-so-naive-about-naive-bayes-58166a6a9eba.

[11] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093-1113, 2014. Available: 10.1016/j.asej.2014.04.011.

[12] Kogan, S., Zeng, Q., Ash, N., & Greenes, R. A. (2001). Problems and challenges in patient information retrieval: a descriptive study. *Proceedings. AMIA Symposium*, 329–333.

[13] Omerhodzic, Ibrahim & Avdakovic, Samir & Nuhanovic, Amir & Dizdarevic, Kemal. (2013). Energy Distribution of EEG Signals: EEG Signal Wavelet-Neural Network Classifier. Int. J. Biol. Life Sci.. 6.

[14] TSAI, C., & WU, J. (2008). *Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Systems with Applications, 34(4), 2639–2649.* doi:10.1016/j.eswa.2007.05.019.

[15] Omerhodzic, Ibrahim & Avdakovic, Samir & Nuhanovic, Amir & Dizdarevic, Kemal. (2013). Energy Distribution of EEG Signals: EEG Signal Wavelet-Neural Network Classifier. Int. J. Biol. Life Sci.. 6.

[16] Widodo, A., & Yang, B.-S. (2007). *Support vector machine in machine condition monitoring and fault diagnosis. Mechanical Systems and Signal Processing, 21(6), 2560–2574.* doi:10.1016/j.ymssp.2006.12.007.

[17] Cai, Y.-D., Feng, K.-Y., Li, Y.-X., & Chou, K.-C. (2003). *Support Vector Machine for predicting α-turn types. Peptides, 24(4), 629–630.* doi:10.1016/s0196-9781(03)00100-1.

[18] Sánchez, J. S., Pla, F., & Ferri, F. J. (1998). *Improving the k-NCN classification rule through heuristic modifications. Pattern Recognition Letters, 19(13), 1165–1170.* doi:10.1016/s0167-8655(98)00108-1.

[19] Gou, J., Yi, Z., Du, L., & Xiong, T. (2012). *A Local Mean-Based k-Nearest Centroid Neighbor Classifier. The Computer Journal, 55(9), 1058–1071.* doi:10.1093/comjnl/bxr131.

[20] Granik, M., & Mesyura, V. (2017). *Fake news detection using naive Bayes classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON).* doi:10.1109/ukrcon.2017.8100379.

[21] Xu, S. (2016). *Bayesian Naïve Bayes classifiers to text classification. Journal of Information Science, 44(1), 48–59.* doi:10.1177/0165551516677946.