Automatic labeling of Twitter data for developing COVID-19 sentiment dataset

by

Sajal Das Shovon

Roll No. 1607030

&

Naimul Hoque Joy

Roll No. 1607099

Department of Computer Science and Engineering

Khulna University of Engineering & Technology

Khulna 9203, Bangladesh

27 March, 2022

Automatic labeling of Twitter data for developing COVID-19 sentiment dataset

by

Sajal Das Shovon

Roll No. 1607030

&

Naimul Hoque Joy

Roll No. 1607099

A thesis submitted in partial fulfillment of the requirements for the degree of

“Bachelor of Science in Computer Science and Engineering”

Supervisor:

K. M. Azharul Hasan

Professor

Department of Computer Science and Engineering

Khulna University of Engineering & Technology

Khulna 9203, Bangladesh --------------------------

Signature

Department of Computer Science and Engineering

Khulna University of Engineering & Technology

Khulna 9203, Bangladesh

27 March, 2022

Acknowledgment

All the praise to the Almighty, whose blessing and mercy succeeded us to complete this thesis work fairly. After that, we humbly acknowledge the valuable suggestions, advice, guidance and sincere co-operation of K. M. Azharul Hasan, Professor, Department of Computer Science and Engineering, Khulna University of Engineering & Technology, under whose supervision this work was carried out. His intellectual advice, encouragement and guidance make us feel confident and scientific research needs much effort in learning and applying and need to have a broad view of problems from different perspectives. We would like to convey our heartiest gratitude to all the faculty members, officials and staff of the Department of Computer Science and Engineering as they have always extended their co-operation to complete this work. Last but not least, we wish to thank our friends and family members for their constant support.

Abstract

The COVID-19 has started expanding through the world and has become a pandemic since January 2020. With the rise of new cases daily along with mass death, nation and society are becoming fearful of it. People from all over the world are expressing their thoughts and views about this pandemic in many social media platforms. Nowadays social media is one of the most common ways to express ideas or verdicts on something. With the improvement of modern computing technology, machines are constantly being conducted to interpret what people express in social media like Twitter, Facebook, Instagram etc. These thoughts or views can be categorized and analyzed based on sentiment. In this research article, we have analyzed the sentiment of people with the help of Twitter data. The tweets are collected as daily raw Twitter data regarding Covid-19 with the help of hydration of tweets using tweet_id from the CrisisNLP website. We have collected daily data between March 1, 2020 - April 30, 2020 and preprocessed twitter text using some preprocessing techniques. In this paper, we have analyzed the sentiment of people what they express in social media by using tweets gathered from Twitter. We have categorized the sentiment of the tweets into five classes namely ’Strongly Negative’, ’Negative’, ’Neutral’, ’Positive’ and ’Strongly Positive’. Initially, we use the Textblob of python for classification. This classification does not show good results and needs massive change as there lie many new terms related to COVID-19 which affects the sentiment of tweets. We have labeled automatically by creating Regular Expression rules with our new corrected word library which is created by analyzing the tweets manually. We have trained a model with the updated labeled dataset with Long short-term memory (LSTM), Logistic Regression (LR), Multinomial Naive Bayes (MNB) and analyzed the performance. We found that our data labeling shows better performance compared to the standard dataset. In this research article, we have introduced three machine learning techniques to analyze the sentiment of Twitter data. They are Logistic Regression, Multinomial Naive Bayes, Dense Neural Network (with hidden layers -Relu) and Dense Neural Network (with hidden layers -Tanh). We have found about 91.7% accuracy in Logistic Regression, 83.5% in Multinomial Naive Bayes and Dense Neural Network shows an accuracy of 96%. Apart from narrating the methods we provide an analogy of three essential machine learning methods in terms of textual analysis and compare their effectiveness. The limitations were also revealed in the wish for better future research. Moreover, we will work on collecting specific regions' data and continue our research with a month-by-month dataset.

Contents

Page

List of Tables

List of Figures

CHAPTER I

Introduction

1.1 Introduction

In the present era of digital communication, social media like Facebook, Twitter, Instagram etc provide the opportunity to share or exchange ideas and information. People can create content to express their feelings and share their opinions about their activities as well as international activities or any social crisis on social media. Approximately, Facebook alone has over 1.2 billion monthly active users, Twitter has over one billion registered users and Instagram has over three hundred million. So, it is easy to understand that these social media are huge platforms to share fun, ideas and information and play a vital role in connecting people and developing relationships throughout the world. In a word, social media creates an opportunity for mass people to share their views and hence connect the people in a single platform.

Sentiment analysis is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Information technology provides profound opportunities to fight infectious disease outbreaks and has a remarkable role, especially in sentiment analysis for social media and this is important due to their tremendous role in analyzing public sentiment. Research shows that many outbreaks and pandemics could have been promptly controlled if experts considered social media data[1]. Sentiment analysis looks at the emotion expressed in a text. It is commonly used to analyze customer feedback, survey responses, and product reviews. Social media monitoring, reputation management, and customer experience are just a few areas that can benefit from sentiment analysis.

1.2 Problem Statement

The rapid growth of COVID-19 creates the demand for understanding the sentiment of mass people in what they express in social media. While some initiatives like medical care, preventative care, there must be an emphasis on social media communications to predict the aspects of COVID-19's effect. There lies enormous data in social media about COVID-19 as it frightens people of all classes. With the advances of NLP, we are able to analyze the methods regarding sentiment analysis. Here textual data are extracted from Twitter in a particular region.

The COVID-19 pandemic created a sensational loss of human life worldwide and presents the challenge to global health, food systems, and the work universe[2]. Social media like Facebook, Twitter, Instagram etc provide the opportunity to share or exchange ideas and information of COVID-19. People express their feelings and share their opinions and their activities about the pandemic and social crisis all over the world [3] [4].

1.3 COVID-19 & Sentiment Analysis

Sentiment analysis is a trending topic these days in many fields like marketing, business models and even in research proposes. As soon as Covid-19 affected the world by storm in 2020, people from all over the world started expressing their thoughts & views on social media. Social media plays a big role in our life nowadays. The public shows their emotions by posting or commenting on various topics like COVID-19 on social media. So it's easy to predict people’s thoughts simply analyze social data like Tweets.

The analysis of public sentiment regarding the COVID-19 vaccine can help government and non-government health organizations to determine and understand public views and mentality about the consequence of COVID-19. With the evolution of machine learning algorithms, it’s easier to analyze people’s sentiments. Sentiment analysis is mainly the classification of opinions contained in tweets. The rapid growth of COVID-19 creates the demand for understanding the sentiment of mass people in what they express in social media.[5][6] Some sentiment examples are like:

Social Data	Sentiment
God, save us from Corona.	Fear
Finally Covid cases are decreasing	Positive
I am having Covid symptoms	Negative

Table 1.1: Some examples of the sentiment of social media data

1.4 Objectives

- The main objective of the thesis is to automatically label the data which are unlabeled now so far. Some words are specially used in the COVID-19 situation. These words have the opposite meaning or are not declared in Textblob. This kind of tweet can't find the real sentiment through Textblob. Besides the new terms (new cases, PCR tests, corona positive/negative) that are related to COVID-19 are not familiar to Textblob. To make this labeling right, we introduce automatic labeling.
- We have used regular expressions to make this type of word meaningful. This will automatically label the given data. After correcting the undefined data we then found the sentiment accurately that was previously wrong.
- The correction against Textblob's labeling is done giving the tweets correct sentiment by creating specific regular expressions regarding COVID-19.
- Finally, we use some machine learning approaches to find the sentiment of Twitter data. We also have found the analogy of our approaches to see which method works better.We compared our work with another standard dataset to measure the correctness of the data labeling.
- This work clarifies public opinion on the COVID-19 pandemic and can guide authorities to overcome needless anxiety during the COVID-19 pandemic.

1.5 Organization of the Thesis

The rest of the thesis is organized as follows:

- Chapter 2 presents short summaries on the related research works on sentiment analysis
- The various methodologies that were pursued along the course of this thesis work are described in detail in Chapter 3.
- Chapter 4 presents the experimental results of our proposed methodologies.
- The limitations, future works and concluding remarks for this thesis are presented in Chapter 5.

CHAPTER II

Literature Review

2.1 Introduction

Various techniques have been employed to successfully perform the detection of the sentiment of the textual data. Sentiment Analysis is an NLP task where a model tries to identify if the given text has positive, negative, or neutral sentiment. Some systems employ the general Natural Language Processing (NLP) technique for detecting the task. Pre-trained NLP models for sentiment analysis are provided by open-source NLP libraries such as BERT, NTLK, Spacy, and Stanford NLP. Some probabilistic models like the Bayesian theorem can be used to be trained with the dataset. Machine Learning-based systems used Logistic Regression, SVM etc to do the job. Deep learning-based systems use CNN, RNN, LSTM and many more approaches. The reviewed literature has thus been described Machine Learning and Deep Learning Approaches.

2.2 Machine Learning and Deep Learning Approaches

Chintalapudi et al. [7] uses deep learning method to detect sentiment using labeled data. Their findings present the high prevalence of keywords and associated terms among Indian tweets during COVID-19. The accuracy of such result depends on the labeled accuracy of dataset. Accuracy for every sentiment was separately calculated. The Bidirectional Encoder Representations from Transformers BERT model produced 89% accuracy and the other three models logistic regression (LR), support vector machines (SVM), and long-short term memory (LSTM) produced 75%, 74.75%, and 65%, respectively.

Manguri et al.[8] used textblob of python for sentiment analysis on twitter data that are collected by COVID-19 and using Twitter API through python library. The data we have collected on twitter are based on two specified hashtag keywords, which are ("COVID-19, coronavirus”). They classified the sentiments in optimistic, neutral and negative sentiment using Naive Bayes method. They also categorized the emotion of tweets for an one week duration based on the polarity. TextBlob library of python Sentiment Analysis techniques applied for collected tweets which are 530232 tweets. The results shown the neutral toll regarding both coronavirus and covid-19 keywords for polarity was significantly high which is more than 50 percent and the large portion of the records were objective which was approximately 64 percent.

Naseem et al. [9] constructed a methodology to classify the sentiment with a large scale Twitter data and find if there lies any misinformation about COVID-19 through the social media. This study also analyzes views concerning COVID-19 by focusing on people who interact and share social media on Twitter. As a platform for their experiments, they present a new large-scale sentiment data set COVIDSENTI, which consists of 90 000 COVID-19-related tweets collected in the early stages of the pandemic, from February to March 2020. The tweets have been labeled into positive, negative, and neutral sentiment classes. They analyzed the collected tweets for sentiment classification using different sets of features and classifiers. Their study contains several ML based algorithm like SVM, Naive Bayes,BERT, and CNN. Their findings show that people had a positive sentiment in the lockdown of February 2020 and by mid-March sentiment is got shifted.

Radaideh et al.[10] introduced a procedure to calculate the sentiment analysis using Naive Bayes and RNN method on a real time scenario. They have collected the data from December 2019 to July 2020 and classified them into positive, neutral and negative sentiment where most of the tweets came out positive. The sentimental analysis found that 630 tweets were positive and people in UAE feels secured, satisfied and internet calling is very useful for them in the prospect of work, education, etc. Only 48 tweets has negative impact because people feel little bit harder in sudden change of culture with in short period of time and 155 tweets has impact that both positive and negative were view and said to be natural. The study found that NB (84%) is more accurate, user friendly and takes less time than RNN (79%) to perform the analysis. Finally, the sentimental analysis reveals that people in UAE were accepting the new culture of internet calling and it is useful for them in the prospect of work and education.

Kaur et al.[11] have analysed sentiment of tweets that are collected based on keyword like coronavirus, deaths, new case etc. In this paper, sentiment analysis was conducted to determine the impact of Twitter data analysis on the mental health status of the people. They have designed an algorithm called Hybrid Heterogeneous Support Vector Machine (H-SVM) and performed the sentiment classification and classified them positive, negative and neutral sentiment scores. They used R programming language for their research. By using RNN and SVM they classified the sentiment in positive, neutral and negative classes. In their research, SVM shows maximum number of tweets as neutral whether RNN describes a maximum number of tweets as positive.

Samuel et al. [12] have proposed a methodology to identify fear sentiment involvement. They have gathered almost 900,000 data using Twitter API and rTweet package in R. They have demonstrated a comparison between fear sentiment and negative sentiment on geo-tagged data too. After researching they have noticed that Naive Bayes has more accuracy than Logistic Regression in terms of both shorter or longer tweets. However they observe that the logistic regression classification method provides a reasonable accuracy of 74% with shorter Tweets, and both methods showed relatively weaker performance for longer Tweets. Finally, they provided a comparison of textual classification mechanisms used in artificial intelligence applications and demonstrated their usefulness for varying lengths of Tweets.

Chakraborty et al. [13] presented the aspects and research of social media analysis in many ways. This paper also deals with social network involved items. This very article indicates many future development of research that are based on social media’s data. This article also addresses the process of capturing data from social media over the years along with the similarity detection based on similar choices of the users in social networks. The techniques of communalizing user data have also been surveyed in this article. This one-of-a-kind paper presents a detailed survey of social networks and its related terms. The works that have been accomplished relating to cluster, community and social networks have been described in its scope. This article mainly aims to bring out the shortfalls of the wide variety of papers making it easy for researches to apply sentiment analysis methods after accumulating data from social media.

CHAPTER III

Automatic labeling of Twitter data & Sentiment Analysis

3.1 Introduction

Automatic labeling of Twitter data for COVID-19 deals with the unlabeled/undefined data that are present in Python library Textblob. Some uncommon words that are not present in library give the wrong or unwanted sentiment result. Moreover, many new buzzing words are come out in time of COVID-19 situation. In terms of analysing sentiment these words don’t give the perfect sentiment. So, at first, we need to label these words/sentences correct label to make the sentiment perfect. After having a right labeled datasets we can predict the sentiment correctly with the help of machine learning approaches.

For training and classifying the dataset in python language, we have used Logistic Regression, Multinomial Naive Bayes and Long short-term memory (LSTM). These are some methods of supervised learning for sentiment analysis. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new input data. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen inputs. We compared it with another COVID-19 [17] to measure our performance.

3.2 Working Procedure

The overall workflow of the system can be divided into two parts: Correcting the sentiment of the wrong labeled data by Textblob & sentiment analysis using the correct labeled dataset. Both of these parts are dependent on machine learning for proper detection and recognition. After correcting the label of the datasets we have trained the model with those. As the dataset now have the correct label, it will train the model more precisely. With clean labeling, the models are trained. At the last stage of the implementation we have the analogy of the models to visualize which model performs better. In the following sections, we provide in-depth discussions on the working principles of our methodology.

The overall working procedure is given below with a diagram:

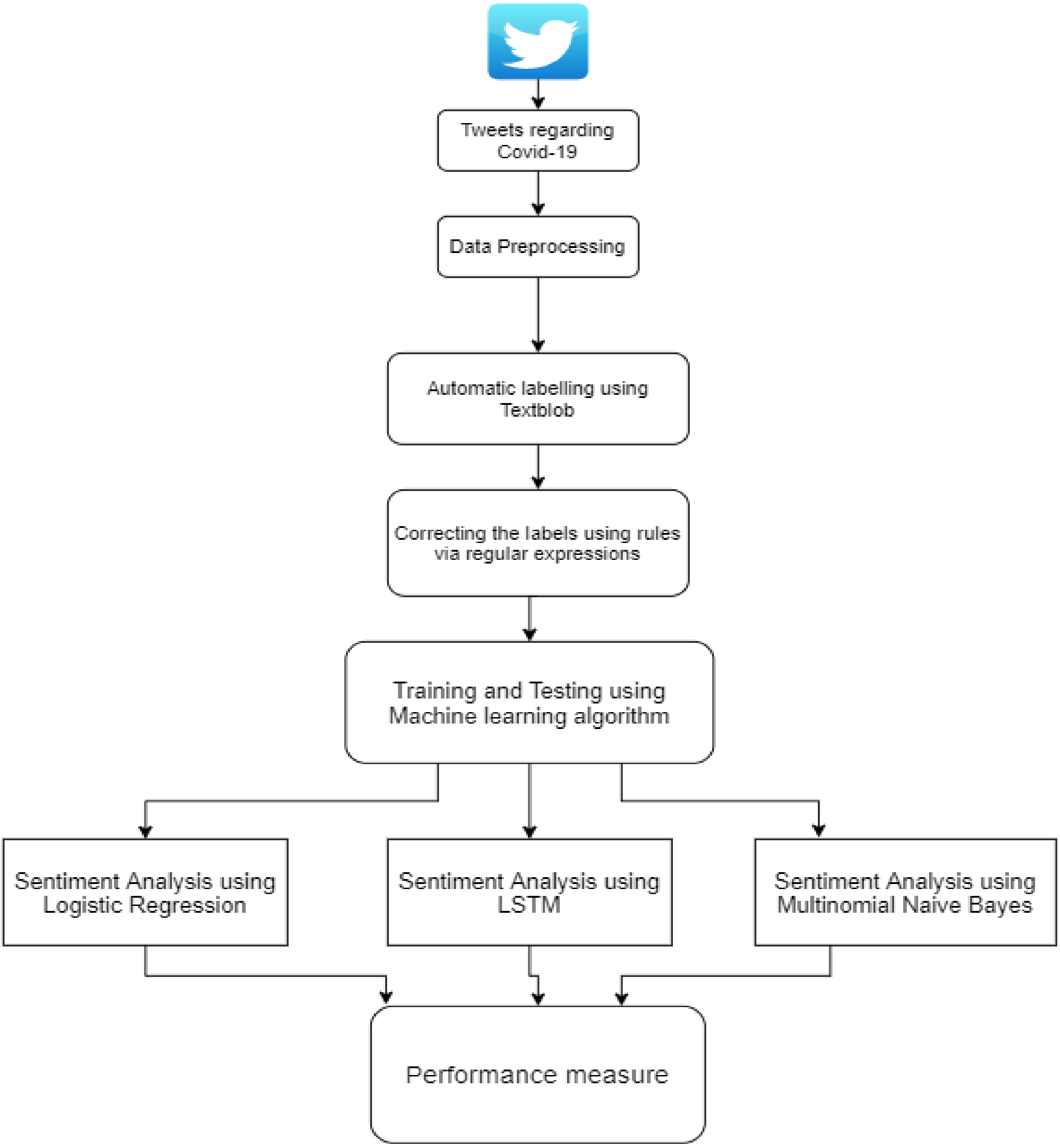
Tweets regarding Covid-19
Data Preprocessing
Automatic labelling using Textblob
Correcting the labels using rules via regular expressions
Training and Testing using Machine learning algorithm
Sentiment Analysis using Logistic Regression
Sentiment Analysis using LSTM
Sentiment Analysis using Multinomial Naive Bayes
Performance measure

Figure 3.1: Flow diagram of proposed methodology

For the procedure of sentiment analysis, firstly we have hydrated twitter data regarding COVID-19. We have collected tweets against twitter id with the help of CrisisNLP[14]. Hence we have made our dataset. We have collected data by twitter extraction & stored our data in a CSV file. The data file consists mainly of data from a definite region. Then we required a preprocessed dataset because unless a clean dataset is made, we can’t train our model properly. So we preprocessed the CSV dataset with some necessary steps. After having the clean dataset we have labeled our processed data using a python library called Textblob. Twxtblob returns the property polarity based on the text. We have done sentiment analysis of tweets and classified them in five multiclasses . Then we have used the labeled dataset to train in Logistic Regression, Multinomial Naïve Bayes & LSTM. This trained model will be helpful to predict the result from a given dataset. We will have a perfect analogy among these three methods and their effectiveness in terms of a particular dataset.

3.2.1 Dataset Acquisition and Description:

Without a clean and processed dataset research on machine learning is difficult. After having the clean dataset we have labeled our processed data using a python library called Textblob and classified. Our target is to generate a dataset containing COVID-19 tweets/sentences. Tweets contains a vast amount of texts that are relatable with COVID-19 based posts/comments from CrisisNLP [14]. This resource consists of Twitter data collected during 19 natural and human-induced disasters. Each dataset contains tweet-ids and human-labeled tweets of the event. [15]. We used a data file consists of almost 1,10,000 tweets. Initially we have used python library twarc to hydrate tweets. Twarc is a toolkit for python to collect twitter data by resembling twitter ID. Twarc provides several different methods for collecting Twitter data. We use twarc dehydrate method to fetch the tweets. It creates a list of tweet ID from a file of tweets. It has a collection of numerous tweets related to COVID-19. This tweets can be fetched even with geographic location.We gathered twitter ID from CrisisNLP and using those IDs we hydrated tweet. The required dataset of tweets is stored in a CSV file.

3.2.2. Data Pre-processing:

For the sentiment analysis we just need some helpful & related keywords to predict the sentiment. So it’s a must to preprocess the dataset and eliminate unnecessary keywords. It will make the dataset clean to analyse the sentiment more precise . A clean dataset is more likely to lead the relevant results. Preprocessing of dataset is a series of steps like removal retweet, extra blank spaces, emojis, punctuation etc. It consists of normalization of text data to improve text matching. There lies some steps to preprocess our dataset as shown below:-

I. Removal of retweet

It means eliminating the retweet of the particular tweet. Retweets are unnecessary as they contain a little relevant sentiment decider word.·

II. Removal of numbers & punctuation

a tweet consists of lots of numbers & punctuation, we must remove them as they hold no important words which can be a part for analysing the sentiment.

III. Stop words removal

In a sentence stopwords are very common & often have repetition. They don’t carry an important meaning. Hence, we can remove stopwords to save computing time and efforts in processing large volumes of text. It is possible to remove stop words using Natural Language Toolkit(nltk).

IV. Stemming

Stemming is a technique that used to extract the base form of the words by removing affixes from them. There are various stemming algorithms in NLTK. We use the SnowballStemmer algorithm to perform this task.

V. Noise removal

Removing unnecessary keywords such as HTML tags, white spaces, emojis or any special character is noted as noise removal. These are not a part of sentiment analysis.

VI. Letter Casing

Converting all letters to either small letter or capital letter. It is mainly done to avoid the duplication of words.

VII. Tokenizing

Turning the whole tweet into tokens is an important step in data preprocessing. Tokenization is the process of splitting the given text into smaller pieces called tokens. Words, numbers, punctuation marks, and others can be considered as tokens. We have used the Natural language toolkit (NLTK) library for tokenization. Tokens are mainly text words that are separated by spaces.

3.3 Data Labeling:

We have used Textblob to label our data. It is done in terms of multiclass sentiment. Sentiment analysis is the process of determining the attitude or emotion of a writer whether it’s positive or negative or neutral. Here we classified our sentiment into 5 classes in this research. They are noted as strongly negative, negative, neutral, positive, strongly positive.

Some unusual words that aren't found in the Textblob library result in an erroneous or unwelcome feeling. Furthermore, numerous new buzzwords have emerged in the context of COVID-19. When it comes to analyzing sentiment, these words aren't quite right. To make the sentiment ideal, we must first categorize these words/sentences correctly.

3.3.1 Initial labeling by textblob:

Sentiment analysis is the process of determining the attitude or emotion of a writer whether it’s positive or negative or neutral. We have used Textblob to label our data. TextBlob is a python library and offers a simple API to access its methods and perform basic NLP tasks. TextBlob is used for initial data labeling.

We use multi class classification (5 classes) namely ’Strongly Negative’, ’Negative’, ’Neutral’, ’Positive’, and ’Strongly Positive’. The sentiment function of Textblob returns two properties, polarity and subjectivity.

Polarity is a float number ranging from [-1,1]. where -1 means strongly negative, -0.5 means negative, 0 means neutral , 0.5 means positive and 1 means strongly positive. Subjective sentences generally indicate personal opinion, emotion or judgment where objective refers to factual information. Subjectivity also lies in a range of [-1,1]. The following table shows the classification criteria according to the polarity.

Polarity Value(P)	**Sentiment**
$-1 \leq P \leq -0.5$	Strongly Negative
$-0.5 \leq P \leq 0$	Negative
0	Neutral
$0 \leq P \leq 0.5$	Positive
$0.5 \leq P \leq 1$	Strongly Positive

Table 3.1: Textblob sentiment classification criteria

3.3.2 Label Correction:

Textblob works as a lexicon-based approach to predict the sentiment from a bag of words. This python library labels the data based on polarity and subjectivity. But in our real time data form twitter, Textblob misplaces some informal languages or words that are unknown to Textblob. These kinds of tweets can’t find the real sentiment through Textblob. Besides the new terms (new cases, PCR tests,corona positive/negative) that are related with COVID- 19 are not familiar to Textblob.

To make this labeling right, we have introduced automatic labeling. It works based on regular expressions that we create from a dataset. We have made the word library and give them the right sentiment. For the time being, we ignore the ”Neutral” sentiment from our dataset as it doesn’t show any sentiment and causes trouble in correctly labeling the data.

Therefore, our revised labeled dataset contains 4 classes instead of 5 namely ”Strongly Negative”, ”Negative”, ”Positive” and ”Strongly Positive”. Table II shows some examples of labeling data.

Regular expression library is used to match the sentiments related to covid-19 with the help of identifier words collected from our dataset. We use these identifiers(shown in Table 3.2) to correctly label the public sentiments. We have used regular expression library (re) and used the findall function (re.findall) to search the specific group of identifiers to to categorize the sentiment correctly for Covid situation.

For example:

- ” My family is tested Covid positive ”, is labeled ” Negative” as it means bad news in COVID-19 situation.
- “I tested Covid negative yesterday”, is labeled “positive” as it means good news in COVID-19 situation.

To do so we develop a word library in order to identify the words hence sentiment correction:

Sentiment Type	Sample Sentiment Identifiers
Strongly Negative	Fake news, new hotspot, Joe Biden syndrome, horrible situation, growing threats, threatened, hate this pandemic, ridicule, worst virus, Oxygen Crisis
Negative	Covid positive, Closing border, community spread, forced quarantine, ban dog meat, politicization of coronavirus can’t afford, high risk countries, complete chaos, coronavirus infection, terrified, total coward, closes churches, new protests trump bashing, panic spreads, homelessness, bigger crisis, touching face, not using sanitizer, coronavirus spreading, shameful comments, sad place, coronavirus vs humanity, coronavirus mess, publishing false stories
Positive	Covid negative, washing hands, won’t get coronavirus, donating fund, shop closed over corona concern, preparing hospitals, hygiene practice, prevent coronavirus, donating, closing borders, well being of people, quarantined, wear surgical mask, self isolation, avoid physical contact, avoid large crowds, Response to coronavirus, response to coronavirus, testing kits, productive meeting, protective equipment
Strongly positive	like a boss, recovered from COVID-19, stand up, love it, What a great idea, stay compassionate, stay safe, good friends, good news, vaccine tested, vaccinated stand up

Table 3.2: Sentiment correction

Example 1: To examine a tweet:

• Tweet: My cousin got Covid positive yesterday.

• TextBlob labeling: Positive.

It’s very clear that this tweet holds a positive sentiment for TextBlob because of the word ’positive’ in the sentence. As it’s a new term regarding to COVID-19 that are not known to Textblob, So we have corrected it as a negative sentiment in our word library and created the regular expression. From now on every tweet that has the word positive’ in it will return a negative sentiment. In the dataset, it is visible that Positive sentiment is greater than negative sentiment. But there is a substantial percentage of negative sentiment in public as COVID-19 has made people concerned, sad and worried.

3.4 Feature Extraction

As our machine can’t deal with text data we need to transform the text into numerical data. For vectorization of text we have used TF-IDF vectorizer. The TF-IDF(Term Frequency- Inverse Document Frequency) implies the importance of a word in a dataset. It creates a TF-IDF matrix that represents the occurrence of a distinct word in our dataset. Bag of words method is used here for simplified representation of words[16].

Term frequency-inverse document frequency combines 2 concepts, Term Frequency (TF) and Document Frequency (DF).

The term frequency is the number of occurrences of a specific term in a document. Term frequency indicates how important a specific term in a document. Term frequency represents every text from the data as a matrix whose rows are the number of documents and columns are the number of distinct terms throughout all documents. Document frequency is the number of documents containing a specific term. Document frequency indicates how common the term is.

Bag of words is a commonly used model in Natural Language Processing. The idea behind this model is the creation of vocabulary that contains the collection of different words, and each word is associated with a count of how it occurs. Later, the vocabulary is used to create d-dimensional feature vectors.

For Example:

D1: Fever, cough, breathing problems are main symptoms of COVID-19

D2: I have fever & cough since yesterday.

Vocabulary could be written as:

V= {fever: 2, cough : 2, breathing : 1, problems : 1, are: 1, main : 1, symptoms : 1, of : 1, COVID_19:1, yesterday:1}

3.5 Logistic Regression

Logistic Regression is a supervised discriminative model. The model builds a regression model to predict the probability that a given data entry belongs to the category. Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. Logistic Regression converts output using the logistic sigmoid function to return a probability value. Logistic regression uses a cost function namely sigmoid function. The sigmoid function is as follows:

$$\phi(t)=\frac{1}{1+e^{-t}} \qquad 3.1$$

But for logistic regression the probability value looks like: P(y = j/Xi) So the probability equation for Ordinal Logistic Regression will be like:

$$P(y \leq j/Xi) = \phi(\theta j\text{-}w^{TXi}) = \frac{1}{1+\exp(w^{TXi}-\theta j)} \qquad 3.2$$

where θ and w are two vectors to be calculated from the dataset. The w vector contains the model learned weights and t values are the feature values.

- A feature representation of the input. For each input observation $x^{(i),}$ this will be a vector of features [x_1, x_2,..., x_n]. We will generally refer to feature i for input $x^{(j)}$ as $x_i^{(j)}$

sometimes simplified as x_i, but we will also see the notation f_i , $f_\mathrm{i}(x)$, or, for multiclass classification, $f_\mathrm{i}(c, x)$.

- A classification function that computes y^, the estimated class, via *p(y|x).*

Logistic regression has mainly two phases:

- training: we train the system (specifically the weights *w* and *b*) using stochastic gradient descent and the cross-entropy loss.
- testing: Given a test example x we compute *p(y|x)* and return the higher probability label $y = 1$ or $y = 0$.

To make a decision on a test instance— after we've learned the weights in training— the classifier first multiplies each x_i by its weight w_i, sums up the weighted features, and adds the bias term *b.* The resulting single number *z* expresses the weighted sum of the evidence for the class.

$$z = (\sum w_i x_i) + b \qquad 3.3$$

Then we'll represent such sums using the dot product notation from linear algebra. The dot product of two vectors a and b, written as a · b is the sum of the products of the corresponding elements of each vector. Thus the following is an equivalent formation to Eq. 5.3:

$$z = w \cdot x + b \qquad 3.4$$

But note that nothing in Eq. 5.4 forces z to be a legal probability, that is, to lie between 0 and 1. In fact, since weights are real-valued, the output might even be negative; z ranges from −∞ to ∞.

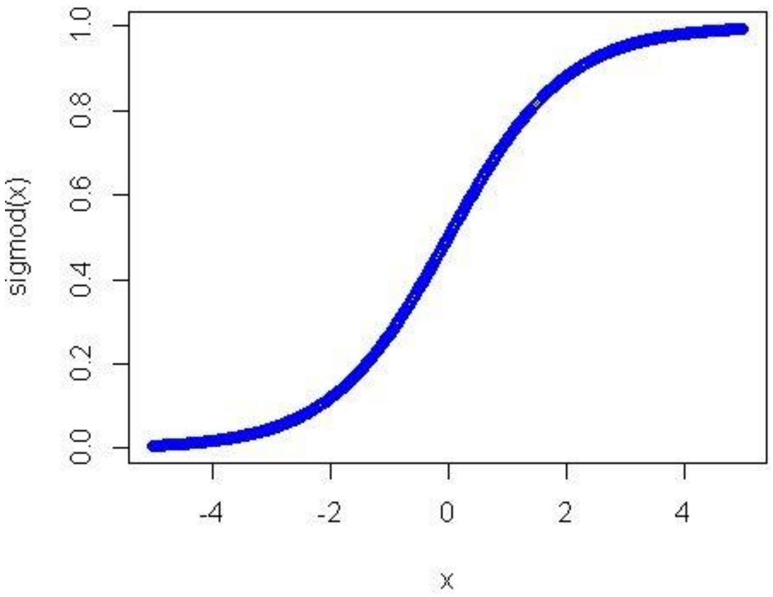

Figure 3.2: The sigmoid function ϕ(t) takes a real value and maps it to the range [0,1].It is nearly linear around 0 but outlier values get squashed toward 0 or 1.

The sigmoid has a number of advantages; it takes a real-valued number and maps it into the range [0,1], which is just what we want for a probability. Because it is nearly linear around 0 but has a sharp slope toward the ends, it tends to squash outlier values toward 0 or 1and it's differentiable, which is handy for learning.

3.6 Multinomial Naive Bayes

Multinomial Naive Bayes is a learning algorithm frequently used in text classification problems. Naive Bayes is based on Bayes' theorem. It is commonly used in text classification, sentiment analysis. It is mainly based on the Bayes theorem which is shown below. It mainly calculates the probabilities of the classes for classifications. There are many types of Naive Bayes Classifiers. As we have 5 classes for sentimental analysis of the COVID-19 tweets, we use here Multinomial Naive Bayes:

$$P(A \mid B) = \frac{P(B|A).P(A)}{P(B)} \qquad 3.5$$

where A and B are two independent classes.

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf also works.

Naive Bayes Summary:

- For each token, it calculates the conditional probability of that token given each class

does this for every token and both classes

- To make a prediction it calculates conditional probability of a class given the token in that message

Let’s see how Multinomial Naive Bayes actually works:

For the computation we have generated a count table from training data like this:

Type	**Positive(W)**	**Neutral(W)**	**Negative(W)**
Positive(S)	200	50	50
Neutral(S)	50	400	50
Negative(S)	50	50	100
Total	**300**	**500**	**200**

Table 3.3: Sentiment calculation for Naïve Bayes

Now from 1000 word we are about to find out a probability. If a word contains a positive word, a neutral word and a negative word what sentiment the MNB will give?

Probability of being positive sentiment,

Y 1(P(S)|P(W), N(W), Ng(W)

$$=\frac{\frac{300}{100}*\frac{200}{300}*\frac{50}{300}*\frac{50}{300}}{0.3+0.5+0.2}$$

$= 5.5 * 10−3$

Similarly we can find probability of any sentences contain any number of positive/negative/neutral word.

3.7 Long Short-Term Memory (LSTM)

LSTM is a deep learning architecture that works with sequential data to solve classification problems. Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems.

In LSTM-

- The cell state act as a transport highway that transfers relative information all the way down the sequence chain. It can be think of it as the “memory” of the network.
- A LSTM unit consists of an LSTM cell, input gate, forget gate and output gate. LSTM can forget the previous values, add new values (new memory) to its data. So even information from the earlier time steps can make it’s way to later time steps, reducing the effects of short-term memory.
- As the cell state goes on its journey, information get’s added or removed to the cell state via gates. The gates are different neural networks that decide which information is allowed on the cell state. The gates can learn what information is relevant to keep or forget during training.

3.7.1 LSTM components:

Sigmoid :

Gates contains sigmoid activations. A sigmoid activation is similar to the tanh activation. Instead of squishing values between -1 and 1, it squishes values between 0 and 1. That is helpful to update or forget data because any number getting multiplied by 0 is 0, causing values to disappears or be “forgotten.” Any number multiplied by 1 is the same value therefore that value stay’s the same or is “kept.” The network can learn which data is not important therefore can be forgotten or which data is important to keep.

Forget gate:

First, we have the forget gate. This gate decides what information should be thrown away or kept. Information from the previous hidden state and information from the current input is passed through the sigmoid function. Values come out between 0 and 1. The closer to 0 means to forget, and the closer to 1 means to keep.

Input Gate*:*

To update the cell state, we have the input gate. First, we pass the previous hidden state and current input into a sigmoid function. That decides which values will be updated by transforming the values to be between 0 and 1. 0 means not important, and 1 means important. You also pass the hidden state and current input into the tanh function to squish values between -1 and 1 to help regulate the network. Then you multiply the tanh output with the sigmoid output. The sigmoid output will decide which information is important to keep from the tanh output.

Cell State:

Now we should have enough information to calculate the cell state. First, the cell state gets pointwise multiplied by the forget vector. This has a possibility of dropping values in the cell state if it gets multiplied by values near 0. Then we take the output from the input gate and do a pointwise addition which updates the cell state to new values that the neural network finds relevant. That gives us our new cell state.

Output Gate*:*

Last we have the output gate. The output gate decides what the next hidden state should be. Remember that the hidden state contains information on previous inputs. The hidden state is also used for predictions. First, we pass the previous hidden state and the current input into a sigmoid function. Then we pass the newly modified cell state to the tanh function. We multiply the tanh output with the sigmoid output to decide what information the hidden state should carry. The output is the hidden state. The new cell state and the new hidden is then carried over to the next time step.

Fig. 3.3 shows the working procedure of LSTM for sentiment detection:

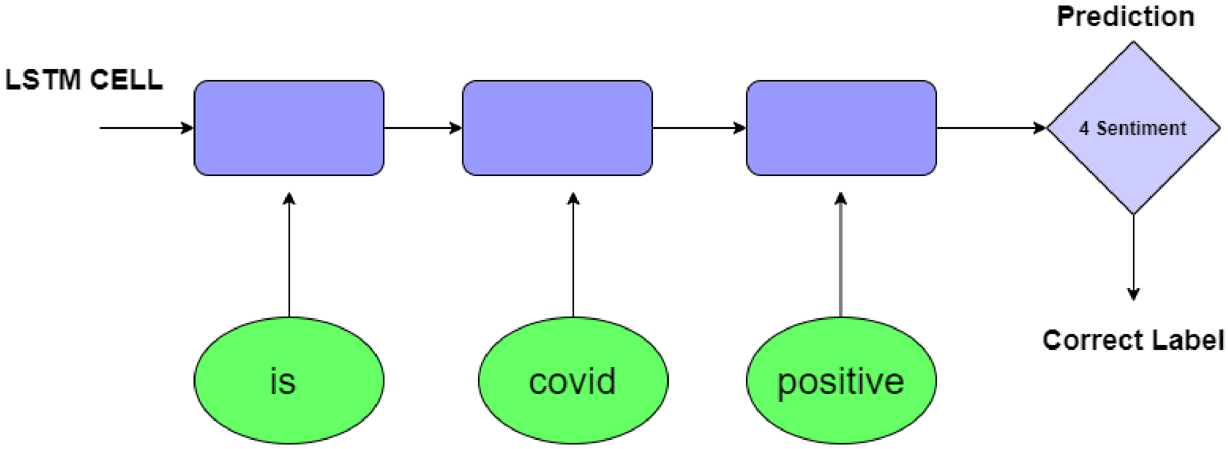

Fig 3.3: Working function of LSTM

There are 1,10,000 tweets in our dataset where no tweets have more than 50 words. Therefore, the maximum text length is limited to 50. We have used the LSTM with 128 hidden units with activation function ’softmax’ to classify the 4 classes. The model is trained for 100 epochs with batch size 32. To prevent the overfitting, we have set the dropout rate to 0.3 with the learning rate of 0.05.

Let’s see some example how our proposed methodology works:

Example:1

- Raw tweet : RT @ConradGoode: The missing six weeks: how Trump failed the biggest test of his life https://t.co/SAIHSF1fdF
- Pre- processed tweet: miss six week trump fail biggest test life
- Label : Strongly Negative.

Example 2 :

- Raw tweet: In an effort to slow the spread of the coronavirus, Governor Gavin New some announces all public schools in the state will remain closed for the rest of the year. I spoke to one high school senior who was really looking forward to a normal graduation.https://t.co/w2t91TUY1a
- Pre processed data: effort slow spread coronavirus governor gavin newsom announc public school state remain close rest year spoke one high school senior realli look forward normal graduat
- Label : Positive

Chapter IV

Experimental Results

At the stage of implementation of this thesis, we have analyzed our experiment results of our dataset1(March 2020) and dataset2(April 2020) on Logistic Regression, Multinomial Naive Bayes and LSTM. We have our visualization in many ways and have the analogy of the models below.

4.1 Experimental Setup

In this chapter, we have done the implementation and related works. The experiments and analysis processes are done on a computer with a core i5 processor having 4 cores with each core having 2.5 GHz Speed. Also the system had 8 GB of Ram. The Graphics card that we have used is the NVIDIA GeForce GTX 1050 ti. The integrated development environment (IDE) that we have used is Anaconda, pycharm, jupyter notebook and kaggle.

4.2 Performance Analysis

This section shows the performance comparison of our data labeling. The dataset contains 1,10,000 tweets. The dataset is available at [21].

The sentiment distribution after our data labeling is shown in Fig. 4.1

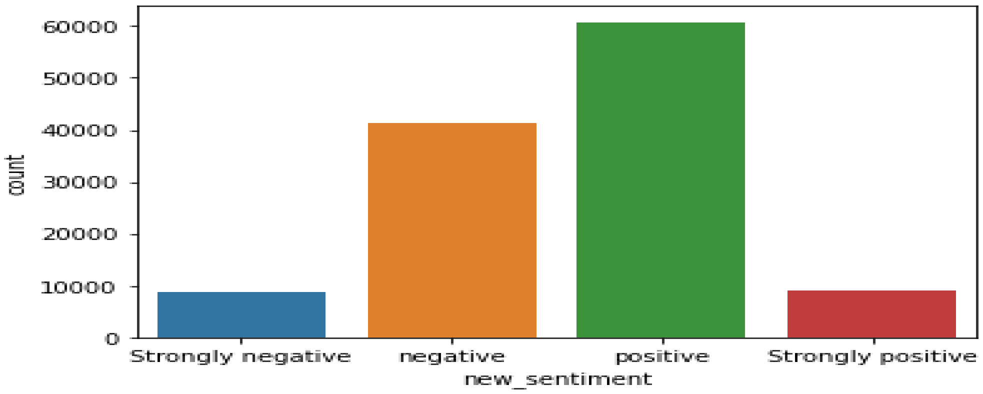

Fig. 4.1: Distribution of sentiments for Covid-19 twitter data

To measure the correctness of the labeling we compare our work with the dataset [20]. It is a Twitter dataset of Indian users during the COVID-19 lockdown period in India. The dataset is used in sentiment detection research in [8]. There are 3090 cleaned tweets in the dataset on the topics coronavirus, lockdown, etc. Fig. 4.2 shows the sentiment distribution of the dataset. In the following we call our developed dataset [21] as dataset1 and dataset from [20] as dataset2.

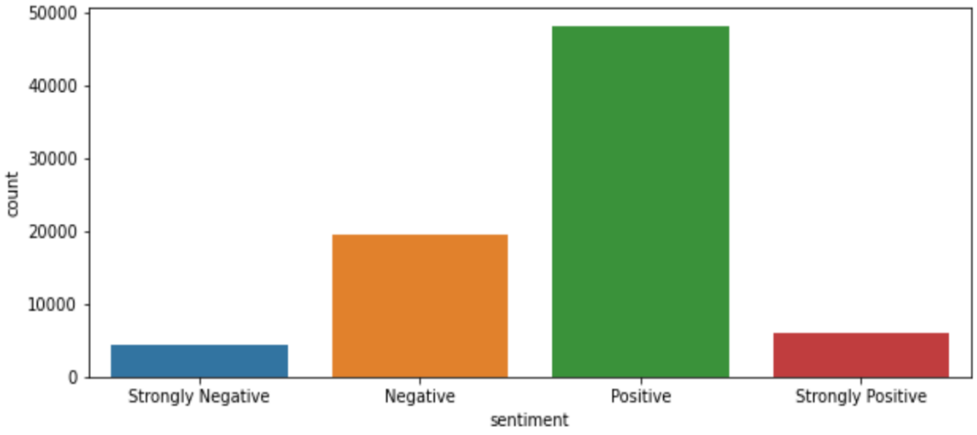

Fig. 4.2: Sentiment distribution of reference data

We have seen a slight decrease of percentage in Strongly negative ,negative,positive data as we increase our dataset. The neutral Data percentage increased almost by 6% and became 60.4% in the 300K dataset which we can see in Fig. 4.3 .These don't provide much information for public sentiment and that’s why we removed these data.

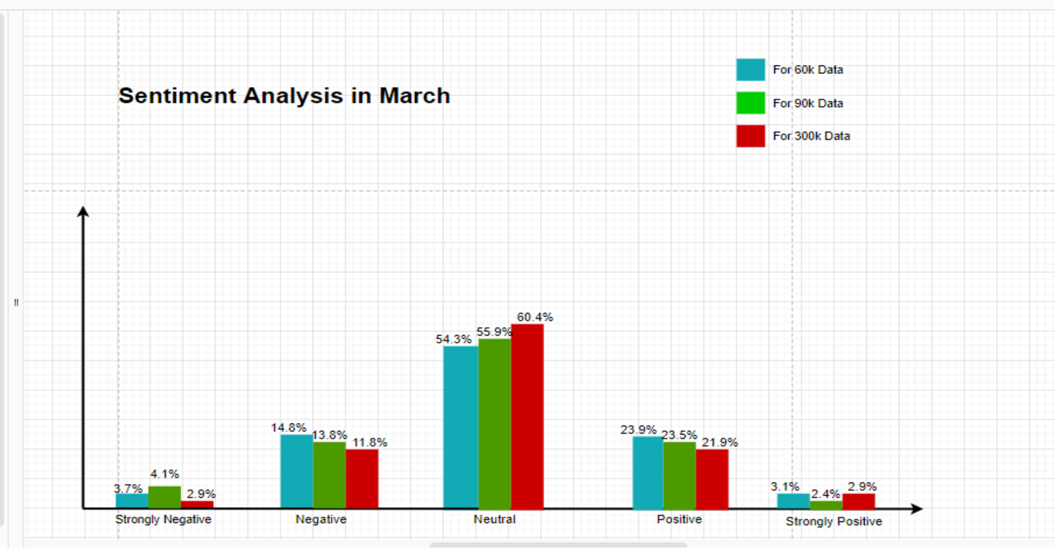

Fig 4.3: March 2020 data sentiment comparison within 60k,90k,300k data

4.3 Result analysis:

Table 5.1 shows the accuracy comparison of dataset1 and dataset2. The LSTM shows better accuracy in comparison with Logistic Regression (LR) and Multinomial Na¨ıve Bayes (MNB). LSTM works better on a large scale dataset. As our dataset is quite large, LSTM gets higher accuracy.

Method	**Dataset 1**	**Dataset 2**
Multinomial Naive Bayes	79%	75%
Logistic Regression	92%	90%
LSTM	96%	95%

Table 4.1: Accuracy graph

Table 4.2 and Table 4.3 show the precision, recall and F1 score for comparison of dataset1 and dataset2. The LSTM shows a better performance than MNB and LR. LSTM works better on a large scale dataset.

As our dataset is quite large, LSTM gets better performance. Adding to this, LR, MNB and LSTM are used as supervised machine learning methods needed. In precision, recall and F1 score LSTM outscores both MNB and LR.

As our dataset varies the data and not perfectly balanced, LSTM provides higher precision and recall. Even the F1 score varies for three models as our dataset is imbalanced. That means LSTM perfectly handled dataset1 and dataset2.

Method	**Precision**	**Recall**	**F1 score**
MNB	0.87	0.62	0.68
LR	0.91	0.89	0.90
LSTM	0.96	0.96	0.96

Table 4.2: Performance measure for Dataset1

Method	**Precision**	**Recall**	**F1 score**
MNB	0.83	0.44	0.44
LR	0.88	0.83	0.85
LSTM	0.94	0.94	0.94

Table 4.3: Performance measure for Dataset2

4.4 Comparison with existing method

The following topic is now very much trending these days. We will now have an analogy between the existing methods and our methodology:

Existing Methods	Dataset Size	Sentiment Class	Used Models	Accuracy
Manguri et al.[8]	500,000 tweets	3 classes (optimistic, neutral and negative)	Naïve Bayes	-
Naseem et al. [9]	90,000 tweets	3 classes (positive, neutral and negative)	SVM, RF, NB, DT	SVM(83.9%), RF(83.6%), NB(76.5%), DT(78.1%) (On a subset of main dataset containing 30,000 tweets)
Radaideh et al.[10]	833 tweets	3 classes (positive, neutral and negative)	Naïve Bayes & RNN	Naïve Bayes(83%) & RNN(79%)

Table 4.4: Performance of existing methods

Where our method is similar or dissimilar in in some extent. Just like:

1. **Dataset:** Primarily we have the dataset size of 300,000. After automatic labeling & discarding the neutral data, the final dataset is 110,000.
2. **Sentiment Class:** At first we classify the dataset in 5 multiclass, called strongly negative, negative, neutral, positive & strongly positive. After discarding the neutral sentiment we have 4 classes finally.
3. **Used Models:** We have used three models called Multinomial Naïve Bayes, Logistic Regression & LSTM to train our dataset.
4. **Accuracy:** Multinomial Naïve Bayes, Logistic Regression & LSTM have the accuracy of 79%, 92%, 96% respectively.

This is the analogy of the existing model & our model. We have he dataset automatically labeled which is not done on the described existing models. Besides 4 multiclass is unique in our model. The accuracy is also higher as we have perfectly labeled dataset.

Chapter V

Conclusion

5.1 Summary:

The twitter dataset was a mixture of different types of sentiment. We show that standard functions such as the textblob of python have some miscalculations for labeling the sentiment. The correction against Textblob’s labeling is done giving the tweets correct sentiment by creating specific regular expressions regarding COVID-19. After correcting the label of the dataset we have performed the sentiment analysis. After performing the analysis we have the results that are noted below:

i) We used the sentiment correction table to automatically correct the Textblob label.

ii) We used Logistic regression, Multinomial Naive Bayes, LSTM to train our data.

iii) It has been found that the LSTM has more accuracy than the Logistic Regression and Multinomial Naive Bayes method.

iv) We found most responses neutral in comparison to positive or negative sentiment.

v) The bigger the dataset the more the neutral sentiment whether both positive & negative sentiment gets lower .

This work clarifies public opinion on COVID-19 pandemic and can guide authorities to overcome needless anxiety during COVID-19 pandemic

5.2 Limitations:

- Our dataset consists of only English tweets. We compelled to discard other language's tweets.
- We collected dataset from Twitter only despite of having other social media.
- We have made word library to correct the Textblob labeling of our dataset with respect to 110k data. So we there is scope for doing better with more dataset.
- We couldn't collect geographical based data.

5.3 Future work:

We tried our best to make the method more successful. But for some limitations there is something that we are about to improve to get better sentiment precision in future. Our future works are as below:-

- To enlarge the dataset and investigate further sentiment analysis and undergo an analogy month by month dataset.
- To compare the sentiment with pre-covid vaccine tweets and post-Covid vaccine tweets to visualize more about sentiment.
- To make more regular expression rules by analyzing more Covid-19 Twitter data regarding public sentiments.

REFERENCES

[1] R. Singh, R. Singh, and A. Bhatia, “Sentiment analysis using machine learning technique to predict outbreaks and epidemics,” Int. J. Adv. Sci. Res, vol. 3, no. 2, pp. 19–24, 2018.

[2] S. Chawla, M. Mittal, M. Chawla, and L. Goyal, “Corona virus-sars-cov-2: an insight to another way of natural disaster,” EAI Endorsed Transactions on Pervasive Health and Technology, vol. 6, no. 22, p. e2, 2020.

[3] M. M. R. Hinnawi, “The role of social media in lifelong informal learning among members of society,” in 2018 JCCO Joint International Conference on ICT in Education and Training, International Conference on Computing in Arabic, and International Conference on Geocomputing (JCCO: TICET-ICCA-GECO). IEEE, 2018, pp. 1–13.

[4] V. Kagan and V. S. Subrahmanian, “Understanding multistage, multi-modal, multimedia events in social media,” in 2018 International Workshop on Social Sensing (SocialSens). IEEE, 2018, pp. 4–4.

[5] S. Vanaja and M. Belwal, “Aspect-level sentiment analysis on e-commerce data,” in 2018 International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2018, pp. 1275–1279.

[6] A. J. Nair, G. Veena, and A. Vinayak, “Comparative study of twitter sentiment on covid-19 tweets,” in 2021 5th International Conference on Computing Methodologies

and Communication (ICCMC). IEEE, 2021, pp. 1773–1778.

[7] N. Chintalapudi, G. Battineni, and F. Amenta, “Sentimental analysis of covid-19 tweets using deep learning models,” Infectious Disease Reports, vol. 13, no. 2, pp. 329–339, 2021.

[8] K. H. Manguri, R. N. Ramadhan, and P. R. M. Amin, “Twitter sentiment analysis on worldwide covid-19 outbreaks,” Kurdistan Journal of Applied Research, pp. 54– 65, 2020.

[9]U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, “Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis,” IEEE Transactions on Computational Social Systems, 2021.

[10] A. Radaideh, F. Dweiri, and M. Obaidat, “A novel approach to predict the real time sentimental analysis by naive bayes & rnn algorithm during the covid pandemic in uae,” in 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI). IEEE, 2020, pp. 1–5.

[11] H. Kaur, S. U. Ahsaan, B. Alankar, and V. Chang, “A proposed sentiment analysis deep learning algorithm for analyzing covid-19 tweets,” Information Systems Frontiers, pp. 1–13, 2021.

[12] J. Samuel, G. Ali, M. Rahman, E. Esawi, Y. Samuel et al., “Covid-19 public sentiment insights and machine learning for tweets classification,” Information, vol. 11, no. 6, p. 314, 2020.

[13] K. Chakraborty, S. Bhattacharyya, and R. Bag, “A survey of sentiment analysis from social media data,” IEEE Transactions on Computational Social Systems, vol. 7, no. 2, pp. 450–464, 2020.

[14] Q. C. R. Institute. Crisisnlp. [Online]. Available: https://crisisnlp.qcri.org/

[15] M. Imran, P. Mitra, and C. Castillo, “Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages,” in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Paris, France: European Language Resources Association (ELRA), may 2016.

[16] Y. Zhang, R. Jin, and Z.-H. Zhou, “Understanding bagof-words model: a statistical framework,” International Journal of Machine Learning and Cybernetics, vol. 1, no. 1-4, pp. 43–52, 2010.

[17] G. Preda. [Online]. Available: https://github.com/gabrielpreda/CoViD-19-tweets

[18] T. Vijay, A. Chawla, B. Dhanka and P. Karmakar, "Sentiment Analysis on COVID-19 Twitter Data," 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 2020, pp. 1-7, doi: 10.1109/ICRAIE51050.2020.9358301.

[19] N. Srivats Athindran, S. Manikandaraj and R. Kamaleshwar, "Comparative Analysis of Customer Sentiments on Competing Brands using Hybrid Model Approach," 2018 3rd International Conference on Inventive Computation Technologies (ICICT), 2018, pp. 348-353, doi: 10.1109/ICICT43934.2018.9034283.

[20] A. J. Nair, V. G and A. Vinayak, "Comparative study of Twitter Sentiment On COVID -19 Tweets," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1773-1778, doi: 10.1109/ICCMC51019.2021.9418320.

[21] L. -C. Cheng and S. -L. Tsai, "Deep Learning for Automated Sentiment Analysis of Social Media," 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2019, pp. 1001-1004, doi: 10.1145/3341161.3344821.

[22] Z. Tariq Soomro, S. H. Waseem Ilyas and U. Yaqub, "Sentiment, Count and Cases: Analysis of Twitter discussions during COVID-19 Pandemic," 2020 7th International Conference on Behavioural and Social Computing (BESC), 2020, pp. 1-4, doi: 10.1109/BESC51023.2020.9348291.

[23] X. Guo and J. Li, "A Novel Twitter Sentiment Analysis Model with Baseline Correlation for Financial Market Prediction with Improved Efficiency," 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2019, pp. 472-477, doi: 10.1109/SNAMS.2019.8931720.

[24] D. Li and J. Qian, "Text sentiment analysis based on long short-term memory," 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI), 2016, pp. 471-475, doi: 10.1109/CCI.2016.7778967.

[25] J. Vijayan, D. A. Siby and G. P. V. Sabeen, "Impact of Covid-19 Pandemic on Recruitment Process and its Sentiment Analysis," 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), 2021, pp. 270-273, doi: 10.1109/ACCESS51619.2021.9563327.

[26] S. Bilal, "A Linguistic System for Predicting Sentiment in Arabic Tweets," 2021 3rd International Conference on Natural Language Processing (ICNLP), 2021, pp. 134-138, doi: 10.1109/ICNLP52887.2021.00028.

[27] F. Rahman et al., "An Annotated Bangla Sentiment Analysis Corpus," 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), 2019, pp. 1-5, doi: 10.1109/ICBSLP47725.2019.201474.