Bachelor Thesis Electrical Engineering

June 2021

Aspect Mining of COVID-19 Outbreak with SVM and Naive Bayes Techniques

By
Akhilandeswari Komara

Department of Mathematics and Natural Sciences
Blekinge Institute of Technology SE–371 79
Karlskrona, Sweden

This thesis is submitted to the Department of Mathematics and Natural Sciences at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Bachelor of Science in Electrical engineering

Contact Information: Author: AKHILANDESWARI KOMARAE-
mail: akko19@student.bth.se
Supervisor: Omsri Kumar Aeddula

University Examiner: Irina Gertsovich

Department of Mathematics and Natural Sciences: Blekinge Institute of Technology
Phone: +46 455 38 50 00 SE–371 79 Karlskrona, Sweden Fax: +46 455 38 50 5

Abstract

The outbreak of COVID-19 is one of the major pandemics faced by the world ever and the World Health Organization (WHO) had declared it as the deadliest virus outbreak in recent times. Due to its incubation period, predicting or identifying the paints hadbecome a tough job and thus, the impact is on a large scale. Most of the countries were affected with Coronavirus since December 2019 and the spread is still counting. Irrespective of the preventive measures being promoted on various media, still the speculations and rumors about this outbreak are peaks, that too particular with the social media platforms like Facebook and Twitter. Millions of posts or tweets are being posted on social media via various apps and due to this, the accuracy of news has become unpredictable, and further, it has increased panic among the people. To overcome these issues, a clear classification or categorization of the posts or tweets should be done to identify the accuracy of the news and this can be done by using the basic sentiment analysis technique of data sciences and machine learning. In this project, Twitter will be considered as the social media platform and the millions of tweets will be analyzed for aspect mining to categorize them into positive, negative, and neutral tweets using the NLP techniques. SVM and Naive Bayes approach of machine learning and this model will be developed.

Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor, Omsri Kumar Aeddula, for the continued support of my thesis study and research, for his patience, motivation, and immense knowledge. His guidance has helped me to study and write this thesis all the time. I am extremely appreciative of what he has offered me.

Contents

Abstract 3

Acknowledgments 4

1. Introduction 8

2. Aspect mining 10

2.1 Proposed system 10

2.2 Modules 11

2.2.1 Data gathering 11

2.2.2 Preprocessing 11

2.2.3 Training and test data 12

2.2.4 Processing 12

2.2.5 Interpretation 12

3. Methodology 14

3.1 Machine learning 14

3.1.1 Supervised learning 14

3.1.2 Unsupervised learning 15

3.1.3 Reinforcement learning 15

3.2 Machine learning algorithms 15

3.2.1 Support vector machine (SVM) 15

3.2.2 Naïve Bayes theorem 16

3.3 Python environment 17

3.3.1 Python 17

3.3.2 NumPy: 17

3.3.3 Pandas 17

3.3.4 Matplotlib 17

3.3.5 Sklearn: 17

4. Implementation 19

4.1 Procedure 19

4.1.1 Dataset 20

4.1.2 Preprocessing of the tweets 20

4.1.3 Feature set extraction 21

4.1.4 Training the machine learning model 21

4.2 Classification algorithms used 22

4.3 Validation 22

5. Results 23

5.1 Preprocessing of data 23

5.1.1 Tokenization of text 24

5.1.2 Stop word removal 24

5.1.3 Punctuation removal from text 25

5.1.4 Lemmatization/Stemming of text 25
5.2 Case study 26
6. Conclusion and future work 33
References 34
Appendix

List of figures:

Figure 1. Flow chart of the technique 11
Figure 2. Pictorial representation of approach 19
Figure 3. Train the machine learning model 21
Figure 4. Input for tokenization of text 24
Figure 5. Output for tokenization of text 24
Figure 6. Input for stop word removal 24
Figure 7. Output for stop word removal 24
Figure 8. Input for punctuation removal from text 25
Figure 9. Output for punctuation removal from text 25
Figure 10. Input for lemmatization of text 25
Figure 11. Output for lemmatization of text 26
Figure 12. The data set used 27
Figure 13. Polarity graph of naive Bayes 28
Figure 14. Polarity graph of SVM 29
Figure 15. Comparison graph of two algorithms 30

Chapter

1. Introduction

Coronavirus disease 2019 (COVID-19) was first detected in Wuhan, China, in December 2019 and has spread worldwide in more than 198 countries [1]. Within a couple of months, this outbreak has upset the world, and uncertainties about the future were for the first time evoked. Nowadays, the age of the internet has changed the way people express their views, opinions. It is now mainly done through blog posts, online forums, product review websites, social media, etc. millions of people are using social network sites like Facebook, Twitter, Google Plus, etc. to express their emotions, opinion, and share views about their daily lives. Through the online communities, interactive media will be obtained where consumers inform and influence others through forums [2]. Social media is generating a large volume of sentiment-rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. Moreover, doctors and people the most affected psychologically by the epidemics are the most likely to talk about it on social networks like Twitter, which have become essential in our daily lives. The analysis of internet users' behavior has shown that they are mostly either seeker of information from the web or social media (Facebook, Twitter, microblogs…) [3]. One of the most famous microblogging sites, Twitter has been one of the major ways for information sharing and self-documentation. As the world is fighting with COVID-19 for the last two months and the majority of the people are under lockdown, the importance of Twitter has increased more than ever. Even in the past, people have been using Twitter to communicate, express and disseminate information related to the crisis, be it cyclones, Ebola, floods, or Zika. Twitter has been one of the platforms for millions to express their emotions regarding different issues. Launched in 2006, the microblogging platform Twitter became one of the most popular microblogging providers [4]. Initially,the users were given 140 characters space to tweet but recently, this has been extended to280 characters. With a total number of 330 million daily active users, Twitter has found an enormous following throughout the world [5]. People use Twitter to share their thoughts, as well as to disseminate information. The rapid sharing of the opinions of the user on Twitter has encouraged the researchers to determine the sentiments on almost everything, including sentiments towards COVID-19 and natural calamities. While there is no single accepted psychological theory of basic human emotions, most of the studiesaccept the theory that a simple positive-negative dichotomy cannot be used to decide electronic copy the human emotions as a whole. On the same lines, it is believed that automatic aspect mining must also implement finely tuned algorithms to detail human emotions [6].

Irrespective of the preventive measures being promoted on various media, still the speculations and rumors about this outbreak are peaks, that too particular with the social media platforms like Facebook and Twitter, tweets from famous research web sites such as online source from a community of data and machine learning practitioners (Kaggle.com) [7]. Millions of posts or tweets are being posted on social media via various apps and due to this, the accuracy of news has become unpredictable, and further, it has increased panic among the people. To overcome these issues, a clear classification or categorization of the posts or tweets should be done to identify the accuracy of the news and this can be done by using the aspect-based sentiment analyses technique of data sciences and machine learning [8]. In this project data, from an online source, a community of data science and machine learning practitioners will be considered as the social media platform and the millions of tweets will be analyzed for aspect mining to categorize the positive, negative and neutral tweets using Natural Language Processing (NLP). SVM and Naïve Bayes approach of machine learning and this model will be developed.

Chapter

2. Aspect mining

2.1 Proposed system

Aspect mining of covid-19 outbreak:

Aspect-based Opinion Mining (ABOM) involves extracting aspects or features of an entity and figuring out opinions about those aspects. It is a method of text classification that has evolved from sentiment analysis and named entity extraction (NER). ABOM is thus a combination of aspect extraction and opinion mining. While opinions about entities are useful, opinions about aspects of those entities are more granular and insightful [9].

This work focuses on aspect opinion mining on covid-19 tweets, this work is proposed with the proposed following objectives.

Objectives

- Review the current COVID-19 data analytics and determine the right choice of datasets to choose.
- To identify positive, negative, and neutral tweets, use aspect mining with NLP techniques and machine learning algorithms such as Naive Bayes and support vector machine (SVM).
- Validate the proposed algorithm.

2.2 Modules

The whole approach is depicted by the following flowchart.

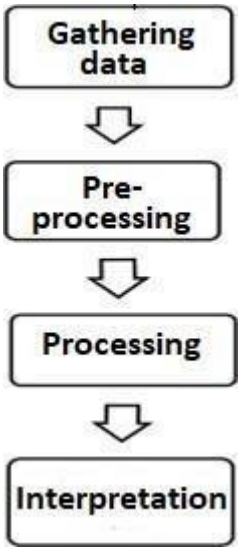


Figure 1. Flow chart of the technique

2.2.1 Data gathering

The dataset contains a collection of data. To assess the current COVID-19 data analytics and determine the right choice of datasets to choose. The dataset is gathered from an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. Tweets will be gathered for all the real-time tweets based on the categorization.

2.2.2 Preprocessing

Preprocessing is a technique that is used to convert raw data into a clean dataset. The data gathered from different sources is in raw format which is not feasible for the analysis. Preprocess with NLP techniques by using NLTK (libraries).

a) Tokenization of text
b) Stop word removal
c) Punctuation removal from the text
d) Lemmatization/stemming

2.2.3 Training and test data

The dataset is divided into testing and training sets. The machine learning model will be trained on the training set, i.e., the model will try to understand the correlations in the training set and then test the model on the test set to examine how accurately it can predict. To assign 80% of the dataset to the training set and therefore the remaining 20% to the test set.

2.2.4 Processing

Classification of data is a two-phase process. In phase one which is called the training phase, a classifier is built using a training set of tuples. The second phase is the classification phase, where the testing set of tuples used for validating the model and the performance of the model is analyzed.

2.2.5 Interpretation

The data set used for further divided into two sets consisting of two-third as a training set and one-third as a testing set. Algorithms applied to naïve Bayes showed the best results. The efficiency of the approaches is compared in terms of accuracy. The accuracy of the prediction model/classifier is defined as the total number of correctly predicted/classified instances.

The implementation process includes:

- **Data processing**

Data processing is the conversion of data into a usable and desired form. This conversion or "processing" is carried out using a predefined sequence of operations either manually or automatically. Most of the processing is done by using computers and thus done automatically.

- **Data cleaning**

Data cleaning is a technique for eliminating noisy data and correcting discrepancies in data. Data cleaning entails applying adjustments to incorrect data. When preparing data for a data warehouse, data cleaning is conducted as a data preparation phase.

- **Data analysis**

Data analysis is the process of cleaning and transforming the collected data. It is a method where once the data is collected, cleaned, and processed it automates the model building process with minimum human effort. During this phase, data analysistools can be used and software that will help us to understand, interpret and derive conclusions based on the requirements.

- **Data visualization**

Data visualization is very common in day-to-day life; it often appears in the form of charts and graphs. In other words, data shown graphically so that it will be easier for the human brain to understand and process it. Data visualization is often used to discover unknown facts and trends. By observing the dataset, a way is found to find out meaningful information.

Chapter

3. Methodology

Data mining:

The method of extracting potentially useful patterns from large data sets is known as data mining. It is a multidisciplinary ability that integrates machine learning, statistics, and artificial intelligence to extract data and determine the likelihood of future events. Data mining insights are used for advertisement, fraud detection, scientific exploration, and other purposes. Data mining is the process of discovering previously unknown but true relationships among data that are hidden, unsuspected, and previously unknown. Knowledge Discovery in Data (KDD), knowledge extraction, pattern analysis, information harvesting, and other terms are used to describe data mining [10].

3.1 Machine learning:

Machine learning is an artificial intelligence (AI) technology that allows systems to learn and develop without having to be specifically programmed. Machine learning is concerned with the development of computer programs that can access data and learn on their own [11]. The learning process starts with observations or data, such as examples, real observation, or instruction, so that trends can be searched in data and make informed decisions in the future based on the examples. The primary objective is for computers to learn on their own, without the need for human interference, and to change their actions accordingly [12].

3.1.1 Supervised learning:

Supervised learning can use branded examples to apply what they have learned in the past to new data in order to predict future events The learning algorithm generates an effective model about the output values based on the analysis of a known training dataset. After enough preparation, the system will provide goals for any new data [13].

3.1.2 Unsupervised learning:

When the information used to train is not confidential or labeled, they are used. Unsupervised learning explores how systems can infer a function from unlabeled data to explain a hidden structure. The system does not find out the right performance, but it explores the data and can infer hidden structures from unlabeled data using datasets [14].

3.1.3 Reinforcement learning:

Reinforcement learning is a type of learning that interacts with its surroundings by performing actions and detecting errors or rewards. The most important characteristics of reinforcement learning are trial and error search and delayed reward. This approach enables machines and software agents to automatically evaluate the best behavior in a given situation in order to optimize their performance [15].

3.2 Machine learning algorithms:

Machine Learning (ML) and the Data Science community are hard at work enhancing epidemiological model forecasts and analyzing information flowing through Twitter for the development of management approaches and assessing the impact of policies to minimize their spread [16]. Various datasets in this regard have been made publicly available. However, more data needs to be captured, developed, and analyzed as the COVID-19 grows worldwide [17].

3.2.1 Support vector machine (SVM):

Support vector machine is a type of machine learning algorithm. Support Vector Machines are typically thought of as a classification method, but they can be used to solve both classification and regression problems. It can handle both numerical and categorical data with ease. To distinguish different groups, SVM creates a hyperplanein multidimensional space. The Support Vector Regression equation is [18]

$$f(x) = x'\beta + b \quad (1)$$

where f(x) is a function, x' and b are integers. The gradient of the line is represented by x', and the y-axis intercept is represented by b (this is sometimes referred to as the vertical intercept).

3.2.2 Naïve Bayes theorem:

Bayes theorem is a formula that describes how to update the probabilities of hypotheses when given evidence. It follows simply from the axioms of conditional probability but can be used to powerfully reason about a wide range of problems involving belief updates [19]

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)} \quad (2)$$

The equation represents the likelihood of event A if event B is real. Event B is often referred to as proof.

P(A) is the value of A (previous likelihood, i.e., the likelihood of an occurrence before the proof is shown). Evidence is the attribute value of an unknown instance (here is event B).

P(B) is a posterior likelihood of B, i.e., the possibility of occurrence after the proof is seen.

P(A|B): the probability of hypothesis A given the data B.

P(B|A): the probability of data B given that hypothesis A was true.

Selection of machine learning algorithms:

Choosing an algorithm is not a trivial decision for every problem. There is no proper algorithm that works for any problem, but in some cases, few algorithms are widely recognized for performing algorithms that are better than others. In this study, machine learning algorithms such as Support Vector Machine and Naïve Bayes algorithms were considered to have been assumed to work well on this problem.

3.3 Python environment

3.3.1 Python:

Python is a powerful programming language that is interpreted, interactive, object-oriented, and general-purpose. Guido van Rossum designed it between 1985 and 1990. Python source code is also available under the GNU General Public License, much like Perl. This tutorial offers ample knowledge of the Python programming language. Python is a programming model that is high-level, interpreted, interactive, and object-oriented. Python is meant to be a very readable language. It frequently uses English keywords instead of punctuation, and it has limited vocabulary structures than other languages [20].

3.3.2 NumPy:

Numerical Python is a library that contains multidimensional array objects as well as a collection of procedures for manipulating those arrays. NumPy helps you to perform mathematical and logical operations on arrays. This tutorial walks you through the fundamentals of NumPy, including its design and climate. It also covers array features, indexing types, and other topics [21].

3.3.3 Pandas:

Pandas is a Python library that offers high-performance, easy-to-use data structures, and statistical software under the BSD license. Python with Pandas is used in a number of academic and business domains, like finance, economics, statistics, analytics, and so on [21].

3.3.4 Matplotlib:

One of the most widely used Python packages for data visualization is Matplotlib. It's a cross-platform library that generates 2D plots from array data. It offers an object-oriented API for embedding plots in Python GUI toolkits such as PyQt applications. It is also compatible with Python and IPython shells, as well as Jupyter notebooks and web application servers [22].

3.3.5 Sklearn:

In Python, Scikit-learn (Sklearn) is the most useful and versatile machine learning library. It uses a Python consistency interface to provide a collection of powerful

methods for machine learning and data mining modeling, such as classification, regression, clustering, and dimensionality reduction. NumPy, SciPy, and Matplotlib are the foundations of this library, which is mostly written in Python [23].

Chapter

4. Implementation

4.1 Procedure:

This approach requires a Twitter data set to analyze it. The data set is collected from a community of data and machine learning practitioners (Kaggle.com). This labeled data set undergoes further different phases of pre-processing which makes it clearer and more understandable than its initial form. This analysis is on labeled data set using the unigram extraction feature technique. This approach proposes a framework where preprocessing techniques are applied to make tweet text which is available in an unstructured format is to understand easily. After preprocessing, required features will be extracted and these features will be formed in the form of feature vectors. Further various machine learning models of supervised nature train the entire data set which is divided in the form of training and testing data set with the support of feature vectors and finally to check accuracy.

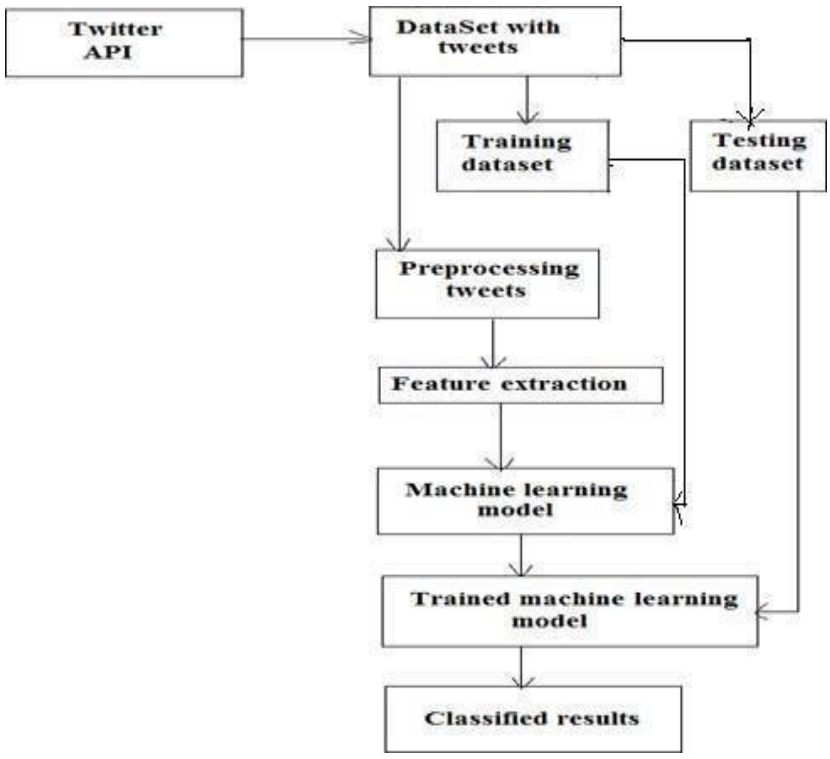


Figure 2. Pictorial representation of the approach

4.1.1 Dataset

This will involve a collection of the data from Twitter. Twitter-related tweet datasets were collected from an online source community of data and machine learning practitioners (Kaggle.com). The downloaded tweets are in the form of a CSV file. By using Python Panda's library, CSV file content will be read and from that CSV file only the tweet text column will be considered, and on this column contains only preprocessing is going to be done in order to get feature set extraction. The tweet data sets can be related to sentiment analysis, opinion mining, product reviews, and this data set is divided into two divisions one is testing data and the other one is training data.

4.1.2 Preprocessing of the tweets

The tweet review text is in the form of an unstructured format from this feature word extraction is difficult. This review text may contain stop words, unwanted characters. Feature words are constructed from this unstructured text only that is the reason preprocessing step is crucial in tweet data analysis. This is a very important step as it decides the extraction of sensitive keywords. It involves syntax-related correction of the tweets as per requirement. This process avoids ambiguity in extracting important features of a data set. Feature word extraction is complex in Twitter text because of misspellings and wrong slang of word usages. To control this, many pre-processing procedure steps are performed before feature extraction. Preprocessing steps involved with the removal of retweets, removal of URLs, if using case sensitive analysis then converting from upper case to lower case or vice versa is required, avoiding misspellings and slang words.

Preprocessing is also involved with the stemming process that replaces words with their root words by reducing no words usage that give the same meanings. This helps in reducing the dimension or size of the feature word set. Few characters do not give meaning such as special characters and digits, thereby it involves the removal of digits and special characters.

4.1.3 Feature set extraction

A feature is an attribute of a keyword of tweet text that can be used as a sensitive word that can help in identifying a problem. The quality and quantity of features are very important as they are important for the results generated by the imposed model. Identification of useful feature words from the tweet data set is feature extraction. These feature word techniques are in the following manner.

Unigram features – only one word is taken at a time and is concluded whether it can identify a feature.

N-gram features – More than one word is taken into consideration for the identification of a feature.

External lexicon – It uses a list of words with predetermined positive or negative emotions.

4.1.4 Training the machine learning model

Machine learning techniques use the entire data set with consideration of a training set and a test set for evaluating a model for the classification of tweets. The training set contains predefined input feature vectors and their corresponding class labels. With the help of this training set, a classification model is developed which tries to classifythe input feature vectors into corresponding class labels. After this, testing data set is used to validate the proposed model by predicting the class labels of hidden feature vectors.

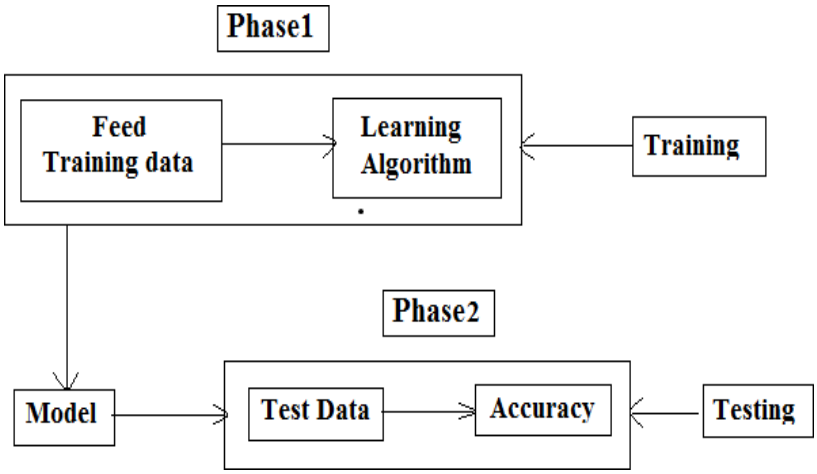


Figure 3. Train the machine learning model

4.2 Classification algorithms used:

Machine learning techniques like Naive Bayes (NB) and Support Vector Machines (SVM) are used to classify tweet posts. Some of the features of a feature vector that can be used for emotion classification are the occurrence of terms, negation, n-grams, and part-of-speech (pos). These features can be used to find out the semantic similar orientationof words, sentences, phrases, and that of data sets. Semantic orientation is the identification of polarity which may be either positive or negative.

4.3 Validation:

K – fold cross-validation method: Cross-validation is a statistical analytic procedure for evaluating the effectiveness of a Machine Learning methodology, as well as a re-sampling approach for validating an algorithm when there is insufficient data. Criteria like ensuring that each fold has the same ratio of outcomes with a given categorical value, such as the class outcome value, can be used to regulate data splitting into folds. K-fold cross-validation is the name of this procedure. The 10-fold cross-validation approach was chosen because it has a lower variance than a single hold-out set estimate approach, which is important when there is a limited amount of data. For diverse datasamples or unique data partitions to build training and test sets, there will be a lot of variation in the results estimate. By comparing more than 10 distinct partitions, the 10- fold cross-validation reduces the performance estimate's sensitivity to data partitioning. To determine the algorithm's average accuracy, the 10-fold cross-validation method will be used. The dataset is divided into 10 folds, with the firstfold being testing data and the subsequent 9 folds being training data. After that, there will be testing data in the second fold, and training data in the remaining nine folds, and so on. An accuracy score for each fold is obtained. It generates ten graphs for each algorithm, each with a different accuracy score. Then the average accuracy based on the accuracy scores of the ten graphs is calculated. This method is applied to both Support Vector Machine and Naïve Bayes Algorithms.

Chapter

5. Results

The research was initiated with the standard and commonly used collection, cleaning of tweets. The process of data preparation, which is outlined briefly below. The tweets have been downloaded using the online community, applying the keyword “Corona”. This has ensured a textual corpus focused on coronavirus, COVID-19, and related phenomena, and reflects the established process for the acquisition of topical data. The raw data was processed and prepared for analysis using python and related packages. In the cleaning process, Python was usedto clean the dataset for further analysis. Since the intention was to use data for academic research, all identifiable abusive words have been replaced with a unique alphanumeric tag word that contained the text "above," but was mixed withnumbers to avoid the use of a set of characters that could have been pre-existing inthe tweets. A complete deletion of abusive words would deprive the data of potential analysis opportunities, and therefore a specially coded algorithm was used.

In addition to the standard use of "Stop words" and cleaning processes, this customized replacement was used. The data set was further evaluated to identify the variables. The blank and irrelevant values have been deleted. The dataset was also further processed based on using “tokenization” which converts text to an analysis of relevant word tokens, "part-of-speech" tags—which tag text artifacts by grammatical category, such as noun or verb, on the needs of each analytical segment of the analysis. "Parsing"—which identifies the underlying textual structure. Elements, "stemming"—which discards prefixes and suffixes by applying rules to create simple basic words, and "lemmatization"—which, like stemming, seeks to transform words into simpler forms and forms. Uses dictionaries and more complex rules and processes than stemming.

5.1 Preprocessing of data

5.1.1 Tokenization of text:

Python tokenization refers essentially to separating a wider body of text into smaller paragraphs, sentences, or even words for non-English language. Various tokenization functions are integrated into the nltk module itself and can be used in programs.

```
import nltk

word_data = "It originated from the idea that there are readers who prefer learn
nltk_tokens = nltk.word_tokenize(word_data)
print (nltk_tokens)
```

Figure 4. Input for tokenization of text

Output:

```
['It', 'originated', 'from', 'the', 'idea', 'that', 'there', 'are', 'readers',
'who', 'prefer', 'learning', 'new', 'skills', 'from', 'the',
'comforts', 'of', 'their', 'drawing', 'rooms']
```

Figure 5. Output for tokenization of text

5.1.2 Stop words removal:

Stop words are English words that do not bring any meaning to a phrase. They can be easily ignored without losing the essence of the sentence. For example, words like that, he, have, etc. Such terms have already been captured in the corpus.

```
from nltk.corpus import stopwords
stopwords.words('english')
print stopwords.words() [620:680]
```

Figure 6. Input for stop word removal

Output:

```
[u'your', u'yours', u'yourself', u'yourselves', u'he', u'him', u'his', u'himsel
u"she's", u'her', u'hers', u'herself', u'it', u"it's", u'its', u'itself', u'the
u'their', u'theirs', u'themselves', u'what', u'which', u'who', u'whom', u'this'
u'that', u"that'll", u'these', u'those', u'am', u'is', u'are', u'was', u'were',
u'been', u'being', u'have', u'has', u'had', u'having', u'do', u'does', u'did',
u'a', u'an', u'the', u'and', u'but', u'if', u'or', u'because', u'as', u'until',
u'while', u'of', u'at']
```

Figure 7. Output for stop word removal

5.1.3 Punctuation removal from the text:

The removal of all punctuation marks from a sentence eliminates all punctuation marks from each phrase.

```
sentence  = "Think and wonder, wonder and think."

tokenizer = nltk.RegexpTokenizer(r"\w+")
new_words = tokenizer.tokenize(sentence)
```

Figure 8. Input for punctuation removal from the text

Output:

```
['Think', 'and', 'wonder', 'wonder', 'and', 'think']
```

Figure 9. Output for punctuation removal from the text

5.1.4 Lemmatization/Stemming of text:

Lemmatization is the method of gathering the various inflected forms of a word so that they can be analyzed as a single object. Lemmatization is equivalent to stemming, but it adds context to the words. So, it connects words of a similar meaning to a single word.

```
# import these modules
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

print("rocks :", lemmatizer.lemmatize("rocks"))
print("corpora :", lemmatizer.lemmatize("corpora"))

# a denotes adjective in "pos"
print("better :", lemmatizer.lemmatize("better", pos ="a"))
```

Figure 10. Input for lemmatization of the text

Output:

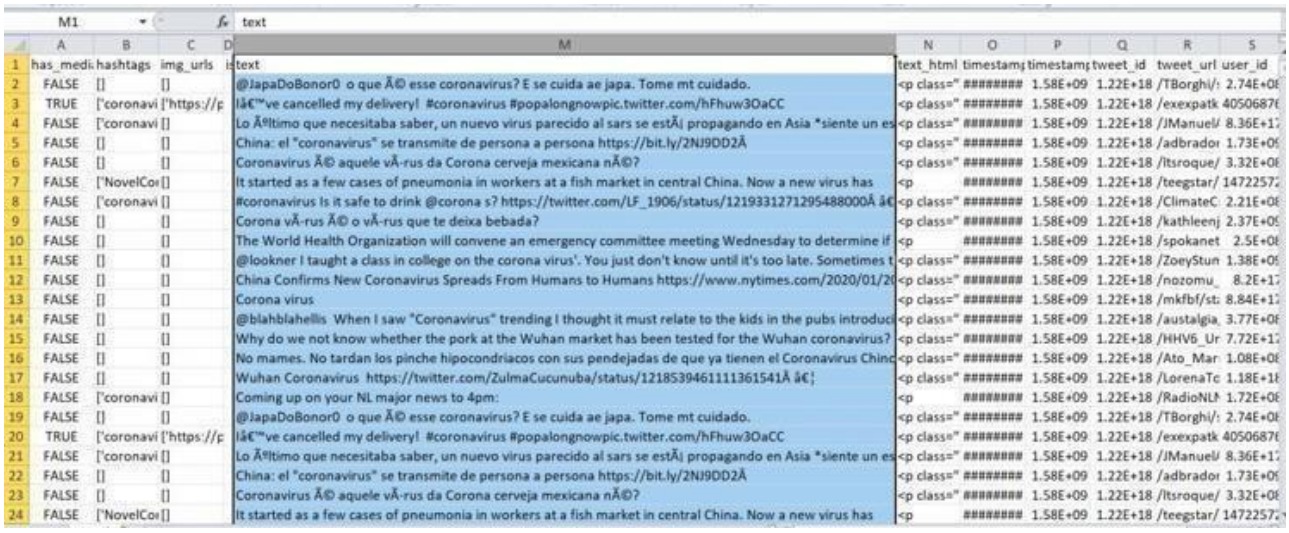
```
rocks : rock
corpora : corpus
better : good
```

Figure 11. Output for lemmatization of the text

5.2 Case study

Kaggle provided a dataset including 10533 COVID-19-related tweets. After that, machine learning techniques were used, and the algorithms were specifically tailored to the tweets dataset in order to uncover the most common patterns in a quick and efficient manner. The categorization results achieved in this work are intriguing, and they suggest that more coronavirus data, as well as new approaches, are needed for further validation and empirical model building. Models thus developed with additional data and methods and using tweet classification methods. The tweet posted on Twitter is in humans understanding language. NLP language helps the machine to understand that tweet and collect the Twitter data from the Kaggle and pre-process with NLP techniques by using NLTK (libraries) i.e., tokenization of text, stopword removal, punctuation removal from text, and lemmatization/stemming.

SVM and Naive Bayes are the algorithms that are used to compare the accuracy of the results. The data is divided into training and testing data. Data has been processed and converted into the desired format. The null values have been removed by cleaning the dataset. The data is divided into 70% training data and 30% test data. By using matplot library, the data visualization has been done with the data for checking the errors. The trained data and test data as the feature extraction be needed. The machinelearning modes like Naïve Bayes and SVM have been used to classify tweet posts. Naïve Bayes generated accuracy 0.9861 and SVM generated accuracy 1.0.



	A	B	C	D	M	N	O	P	Q	R	S
1	has_medi	hashtags	img_urls	is	text	text_html	timestamp	timestamp	tweet_id	tweet_url	user_id
2	FALSE	[]	[]		@JapaDoBonor0 o que Ã© esse coronavirus? E se cuida ae japa. Tome mt cuidado.	<p class="	########	1.58E+09	1.22E+18	/TBorghi/:	2.74E+0
3	TRUE	['coronavi	['https://p		Iâ€™ve cancelled my delivery! #coronavirus #popalongnowpic.twitter.com/hFhuw3OaCC	<p class="	########	1.58E+09	1.22E+18	/exexpatk	4050687
4	FALSE	['coronavi	[]		Lo Ãºltimo que necesitaba saber, un nuevo virus parecido al sars se estÃ¡ propagando en Asia *siente un es	<p class="	########	1.58E+09	1.22E+18	/JManuel/	8.36E+1
5	FALSE	[]	[]		China: el "coronavirus" se transmite de persona a persona https://bit.ly/2NJ9DD2Â	<p class="	########	1.58E+09	1.22E+18	/adbrador	1.73E+0
6	FALSE	[]	[]		Coronavirus Ã© aquele vÃ-rus da Corona cerveja mexicana nÃ©?	<p class="	########	1.58E+09	1.22E+18	/ltsroque/	3.32E+0
7	FALSE	['NovelCor	[]		It started as a few cases of pneumonia in workers at a fish market in central China. Now a new virus has	<p	########	1.58E+09	1.22E+18	/teegstar/	1472257
8	FALSE	['coronavi	[]		#coronavirus Is it safe to drink @corona s? https://twitter.com/LF_1906/status/1219331271295488000Â â€	<p class="	########	1.58E+09	1.22E+18	/ClimateC	2.21E+0
9	FALSE	[]	[]		Corona vÃ-rus Ã© o vÃ-rus que te deixa bebada?	<p class="	########	1.58E+09	1.22E+18	/kathleenj	2.37E+0
10	FALSE	[]	[]		The World Health Organization will convene an emergency committee meeting Wednesday to determine if	<p	########	1.58E+09	1.22E+18	/spokanet	2.5E+0
11	FALSE	[]	[]		@lookner I taught a class in college on the corona virus'. You just don't know until it's too late. Sometimes t	<p class="	########	1.58E+09	1.22E+18	/ZoeyStun	1.38E+0
12	FALSE	[]	[]		China Confirms New Coronavirus Spreads From Humans to Humans https://www.nytimes.com/2020/01/2	<p class="	########	1.58E+09	1.22E+18	/nozomu_	8.2E+1
13	FALSE	[]	[]		Corona virus	<p class="	########	1.58E+09	1.22E+18	/mkfbf/st	8.84E+1
14	FALSE	[]	[]		@blahblahellis When I saw "Coronavirus" trending I thought it must relate to the kids in the pubs introduci	<p class="	########	1.58E+09	1.22E+18	/austalgia	3.77E+0
15	FALSE	[]	[]		Why do we not know whether the pork at the Wuhan market has been tested for the Wuhan coronavirus?	<p class="	########	1.58E+09	1.22E+18	/HHV6_Ur	7.72E+1
16	FALSE	[]	[]		No mames. No tardan los pinche hipocondriacos con sus pendejadas de que ya tienen el Coronavirus Chinc	<p class="	########	1.58E+09	1.22E+18	/Ato_Mar	1.08E+0
17	FALSE	[]	[]		Wuhan Coronavirus https://twitter.com/ZulmaCucunuba/status/1218539461111361541Â â€¦	<p class="	########	1.58E+09	1.22E+18	/LorenaTc	1.18E+1
18	FALSE	['coronavi	[]		Coming up on your NL major news to 4pm:	<p	########	1.58E+09	1.22E+18	/RadioNL	1.72E+0
19	FALSE	[]	[]		@JapaDoBonor0 o que Ã© esse coronavirus? E se cuida ae japa. Tome mt cuidado.	<p class="	########	1.58E+09	1.22E+18	/TBorghi/:	2.74E+0
20	TRUE	['coronavi	['https://p		Iâ€™ve cancelled my delivery! #coronavirus #popalongnowpic.twitter.com/hFhuw3OaCC	<p class="	########	1.58E+09	1.22E+18	/exexpatk	4050687
21	FALSE	['coronavi	[]		Lo Ãºltimo que necesitaba saber, un nuevo virus parecido al sars se estÃ¡ propagando en Asia *siente un es	<p class="	########	1.58E+09	1.22E+18	/JManuel/	8.36E+1
22	FALSE	[]	[]		China: el "coronavirus" se transmite de persona a persona https://bit.ly/2NJ9DD2Â	<p class="	########	1.58E+09	1.22E+18	/adbrador	1.73E+0
23	FALSE	[]	[]		Coronavirus Ã© aquele vÃ-rus da Corona cerveja mexicana nÃ©?	<p class="	########	1.58E+09	1.22E+18	/ltsroque/	3.32E+0
24	FALSE	['NovelCor	[]		It started as a few cases of pneumonia in workers at a fish market in central China. Now a new virus has	<p	########	1.58E+09	1.22E+18	/teegstar/	1472257

Figure 12. Data set used

Naive Bayes algorithm is available in the sci-kit library and sklearn library. In that library sklearn import metrics. Based on metrics the accuracy will be calculated in metrics. accuracy_score (x_test, y_pred)

In a similar way SVM

$$p(h|D) = \frac{p(D|h)\,p(h)}{p(D)} \tag{3}$$

P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.
P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.
P(h|D): the probability of hypothesis h given the data D.
P(D|h): the probability of data D given that the hypothesis h was true.

If the accuracy is 80 percent, our model is underfitting, if it is between 80-100 percent, then best fit, if it is more than 100 percent, then the model is overfitting. In the results, itwas clearly shown that in both the classifiers, with an accuracy of almost 98 percent, our model is ideally suited.

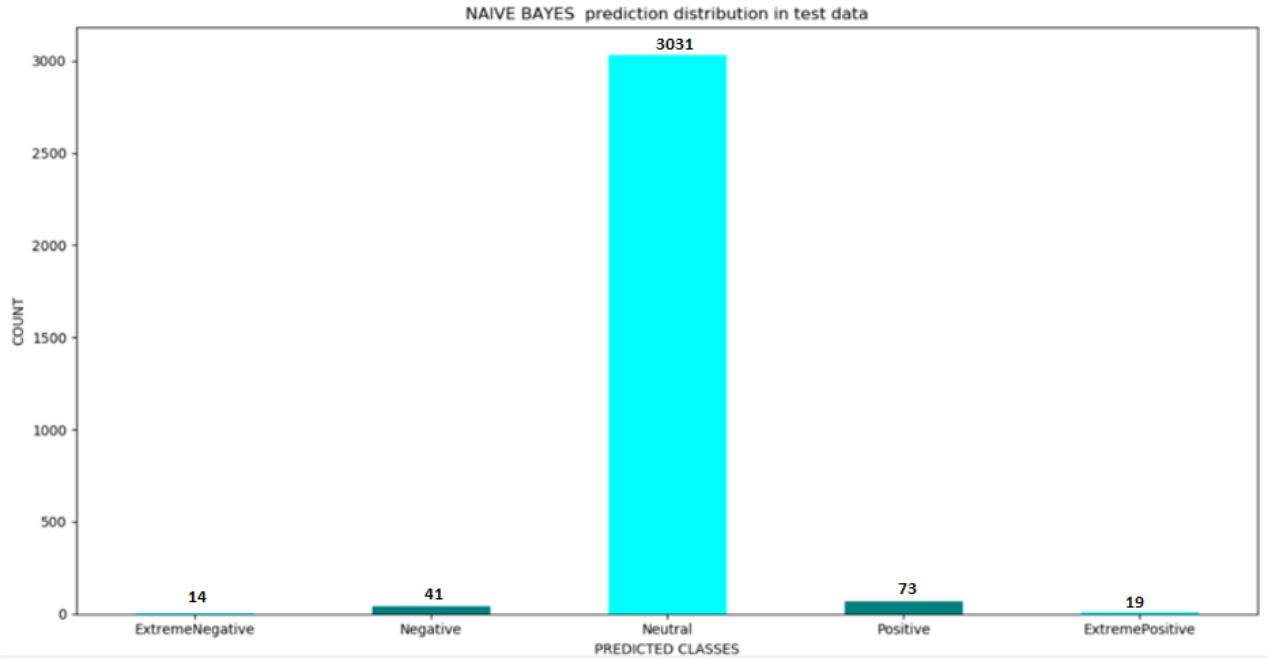


Figure 13. Polarity graph of Naive Bayes

After testing 30 percent of test data, it is observed that the majority of the tweets count isneutral for the Naïve Bayes algorithm.

The count of extreme negative tweets is: 14
The count of negative tweets is: 41
The count of neutral tweets is: 3031
The count of positive tweets is: 73
The count of extreme positive tweets is: 19

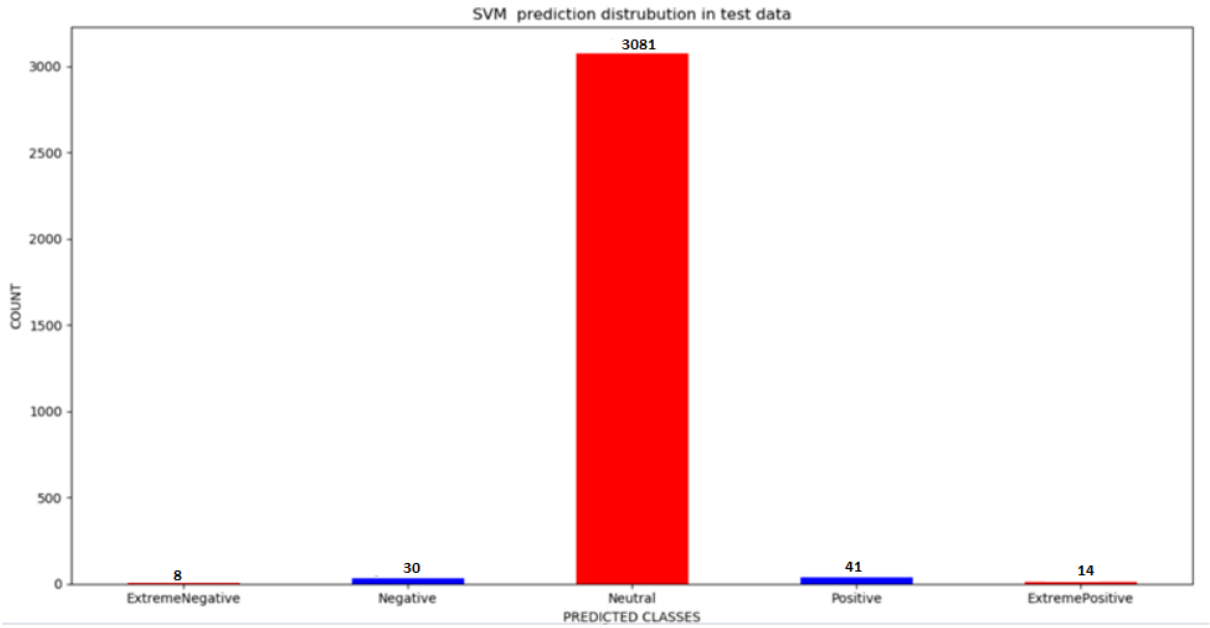


Figure 14. Polarity graph of SVM

After testing 30 percent of test data, it is observed that the majority of the tweets count is neutral for the SVM algorithm.

The count of extreme negative tweets is: 8
The count of negative tweets is: 30
The count of neutral tweets is: 3081
The count of positive tweets is: 41
The count of extreme positive tweets is: 14

In figures 13 and 14 the y- axis is a count of the tweets from the dataset and in the x-axis, the five predicted classes are extreme negative, negative, neutral, positive, extremely positive. This extreme negative tweet is nothing but a tweet that containsmore negative information than positive. If a tweet contains more negative information, then it is a negative tweet. If a tweet contains half negative and half positive, then it is a neutral tweet. If a tweet contains more positive information, then it is a positive tweet. Ifa tweet contains more positive information, than negative then it is an extremely positive tweet.

By comparing graphs with two algorithms:

To validate the proposed algorithm, the 10-fold cross-validation method is used to find the average accuracy of the algorithm. In this 10-fold cross-validation, the dataset will be divided into 10 folds i.e., the first fold is testing data and the remaining 9 parts will be training data. After that second fold will be testing data and the remaining 9 parts will be training data and so on. For every fold, an accuracy score will be obtained. For each algorithm,10 graphs with a different accuracy score will be obtained. The average accuracy based on the 10 graphs accuracy scores will be calculated. For better understanding, graphs are placed in the appendix.

Accuracy score for Naïve Bayes for 10 folds:

Folds	Accuracy
Fold 1	0.98
Fold 2	0.98
Fold 3	0.98
Fold 4	0.98
Fold 5	0.98
Fold 6	0.99
Fold 7	0.99
Fold 8	0.98
Fold 9	0.98
Fold 10	0.98

Avg accuracy score is: 0.98

The average accuracy Score for Naïve Bayes is: 0.98

Accuracy score for Support Vector Machine (SVM) for 10 folds:

Folds	Accuracy
Fold 1	1.0
Fold 2	1.0
Fold 3	1.0
Fold 4	1.0
Fold 5	1.0
Fold 6	1.0
Fold 7	1.0
Fold 8	1.0
Fold 9	1.0
Fold 10	1.0

Avg accuracy score is 1.0

The average accuracy of SVM is 1.0

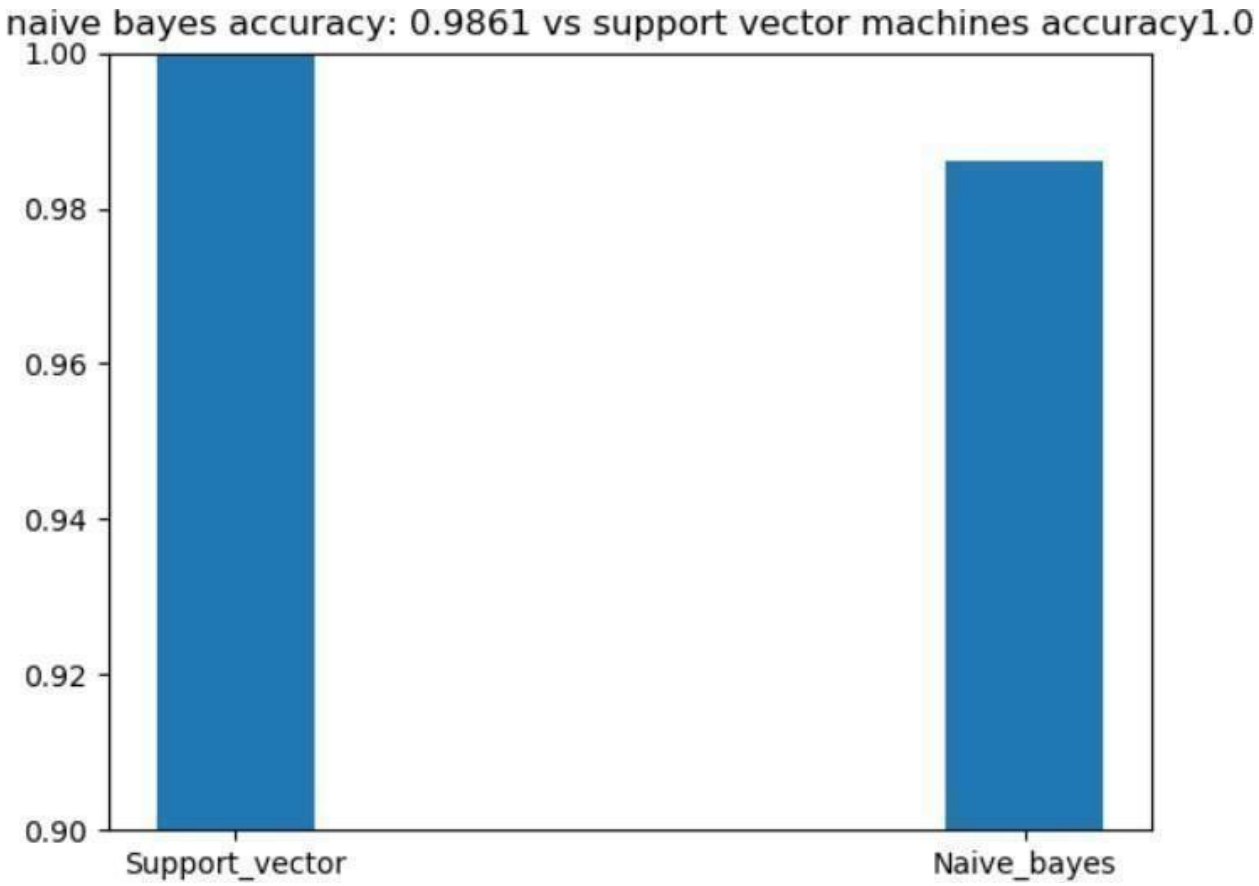


Figure 15. Comparison graph of two algorithms

The above graph shows a comparison between the two algorithms. The first one is the SVM and the second one is Naïve Bayes. The x-axis represents the type of the algorithm and the y-axis represents the accuracy percentage on the test data. Naïve Bayes generates an accuracy of 0.9861 and SVM generates an accuracy of 1.0. The difference between the accuracy of the two algorithms is 0.1. By comparing both algorithms, Support Vector Machine accuracy has more accuracy than Naive Bayes.

Chapter

6. Conclusions and future scope

In conclusion, the present descriptive, exploratory analysis of COVID-19 offers important new information to the international community on the epidemic in China. Twitter was selected to undertake the study because of the credibility of its users and the ease of extracting the tweets. As a first contribution, a dataset of 10533 tweets was extracted using Kaggle.com and preprocessed to highlight useful insights. Machine learning technologies were afterward exploited and the SVM and Naive Bayes algorithms were specially adapted to the tweets dataset to discover the most frequent patterns effectively and efficiently. This project also identifies the difference between SVM and Naïve Bayes algorithms, the SVM algorithm gives accurate results. The derived sentiment highlights our understanding of the tweeters communicating on the COVID-19. It was concluded that people are tweeting words like Pandemic, Death, Quarantine, Hope, Stay Safe, Government, Political, Fight, and Masks with different emotions. Research results show that machine learning methods, such as SVM and Naïve Bayes have the highest accuracy and can be regarded as the baseline learning methods.

The main objective of this work is to analyze COVID-19 related tweets to get the emotion from user's tweets. In this work, it is observed that emotion with machine learning algorithms such as Naïve Bayes and Support Vector Machines. This machine learning approach is statistical and deep learning is neural network-based. This work can be extended in the future to analyze covid related tweet's emotions with deep learning techniques such as artificial neural networks (ANN).

References

[1] M. Ciotti *et al.*, "COVID-19 Outbreak: An Overview," *Chemotherapy*, vol. 64, no. 5–6, pp. 215–223, 2019, doi: 10.1159/000507423.

[2] Md. S. Satu *et al.*, "TClustVID: A novel machine learning classification model toinvestigate topics and sentiment in COVID-19 tweets," *Knowl.-Based Syst.*, vol. 226, p.107126, Aug. 2021, doi: 10.1016/j.knosys.2021.107126.

[3] M. Cinelli *et al.*, "The COVID-19 social media infodemic," *Sci. Rep.*, vol. 10, no. 1, Art. no. 1, Oct. 2020, doi: 10.1038/s41598-020-73510-5.

[4] S. Kushwaha *et al.*, "Significant applications of machine learning for covid-19 pandemic," *J. Ind. Integr. Manag.*, vol. 5, no. 4, Dec. 2020, doi: 10.1142/S2424862220500268.

[5] "COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms | medRxiv." https://www.medrxiv.org/content/10.1101/2020.04.08.20057679v2 (accessed Jun. 04, 2021).

[6] H. Jelodar, Y. Wang, R. Orji, and H. Huang, "Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach," *ArXiv200411695 Cs*, Apr. 2020, Accessed: Aug. 17, 2020. [Online]. Available: http://arxiv.org/abs/2004.11695

[7] "Fusing Logical Relationship Information of Text in Neural Network for Text Classification." https://www.hindawi.com/journals/mpe/2020/5426795/ (accessed Jun. 04, 2021).

[8] P. H. Santoso, F. Fauziah, and N. Nurhayati, "Application of Data Mining Classification for Covid-19 Infected Status Using Algortima Naïve Method: Application of Data Mining Classification for Covid-19 Infected Status Using Algortima Naïve Method," *J. Mantik*, vol. 4, no. 1, Art. no. 1, May 2020, Accessed: Jun. 04, 2021. [Online]. Available: http://iocscience.org/ejournal/index.php/mantik/article/view/740

[9] N. Karisani and P. Karisani, "Mining Coronavirus (COVID-19) Posts in Social Media," *ArXiv200406778 Cs Stat*, Mar. 2020, Accessed: Aug. 12, 2020. [Online]. Available: http://arxiv.org/abs/2004.06778

[10] M. Krause-Traudes, S. Scheider, S. Rüping, and H. Meßner, "Spatial data mining for retail sales forecasting," p. 11, 2008.

[11] S. Marsland, *Machine Learning: An Algorithmic Perspective, Second Edition*. CRC Press, 2015.

[12] J. W. Shavlik, T. Dietterich, and T. G. Dietterich, *Readings in Machine Learning*. Morgan Kaufmann, 1990.

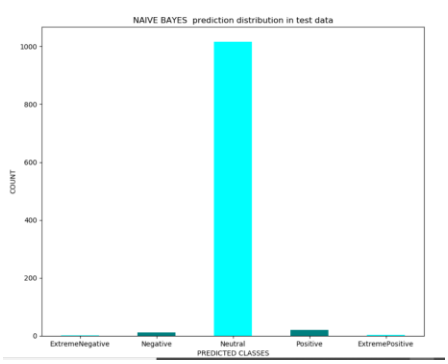
[13] M. Riedmiller, "Advanced supervised learning in multi-layer perceptrons — From backpropagation to adaptive learning algorithms," *Comput. Stand. Interfaces*, vol.

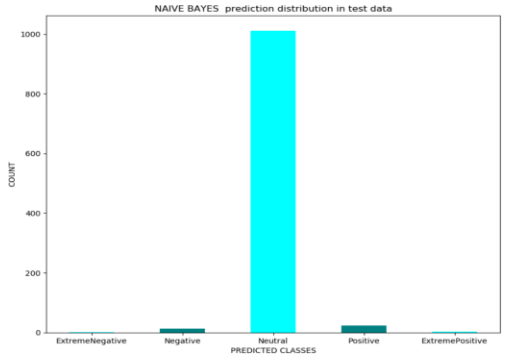
16, no. 3, pp. 265–278, Jul. 1994, doi: 10.1016/0920-5489(94)90017-5.
[14] A. W. Tsui, Y.-H. Chuang, and H.-H. Chu, “Unsupervised Learning for Solving RSS Hardware Variance Problem in WiFi Localization,” *Mob. Netw. Appl.*, vol. 14, no. 5, pp. 677–691, Oct. 2009, doi: 10.1007/s11036-008-0139-0.
[15] R. S. Sutton and A. G. Barto, *Reinforcement Learning, second edition: An Introduction*. MIT Press, 2018.
[16] B. M. Pavlyshenko, “Machine-Learning Models for Sales Time Series Forecasting,” *Data*, vol. 4, no. 1, Art. no. 1, Mar. 2019, doi: 10.3390/data4010015.
[17] M. Mousavizadegan and H. Mohabatkar, “An Evaluation on Different Machine Learning Algorithms for Classification and Prediction of Antifungal Peptides,” *Med. Chem.*, vol. 12, no. 8, pp. 795–800, Dec. 2016.
[18] A. Widodo and B.-S. Yang, “Support vector machine in machine condition monitoring and fault diagnosis,” *Mech. Syst. Signal Process.*, vol. 21, no. 6, pp. 2560–2574, Aug. 2007, doi: 10.1016/j.ymssp.2006.12.007.
[19] S. Durukan and F. Karadagli, “Physical characteristics, fiber compositions, and tensile properties of nonwoven wipes and toilet papers in relevance to what is flushable,” *Sci. Total Environ.*, vol. 697, p. 134135, Dec. 2019, doi: 10.1016/j.scitotenv.2019.134135.
[20] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Mach. Learn. PYTHON*, p. 6.
[21] “Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython - Wes McKinney - Google Books.” https://books.google.se/books?hl=en&lr=&id=v3n4_AK8vu0C&oi=fnd&pg=PR3&dq=numpy+python&ots=rgML9oxzpx&sig=imCQgSpnTzTXbj-aO03R7Mi2Pbo&redir_esc=y#v=onepage&q=numpy%20python&f=false (accessed Aug. 17, 2020).
[22] “Persistence as an aspect | Proceedings of the 2nd international conference on Aspect-oriented software development.” https://dl.acm.org/doi/abs/10.1145/643603.643616 (accessed Feb. 19, 2021).
[23] A. Abraham *et al.*, “Machine learning for neuroimaging with scikit-learn,” *Front. Neuroinformatics*, vol. 8, 2014, doi: 10.3389/fninf.2014.00014.

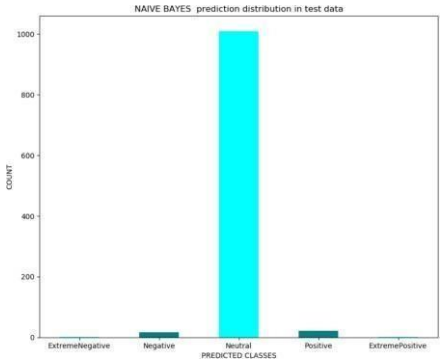
Appendix:

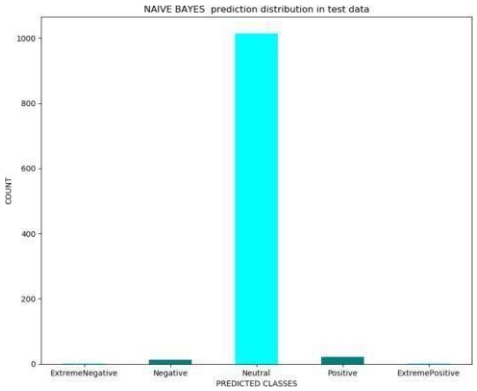
The below graphs show the accuracies of the 10 folds for Naïve Bayes and the values in the results section during validation are taken from these graphs.

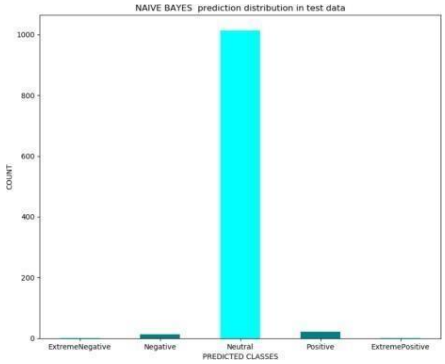
Accuracy graph of Naïve bayes in each fold:

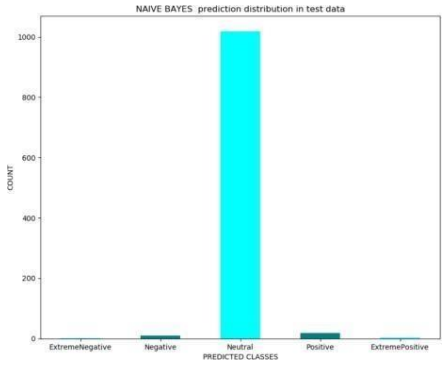


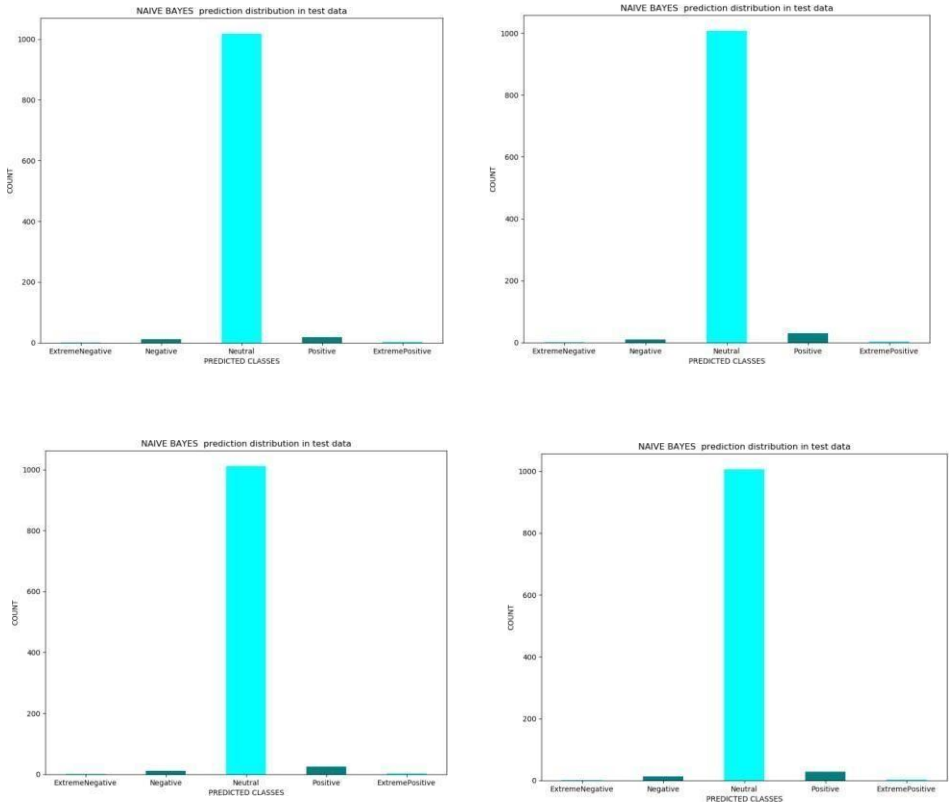






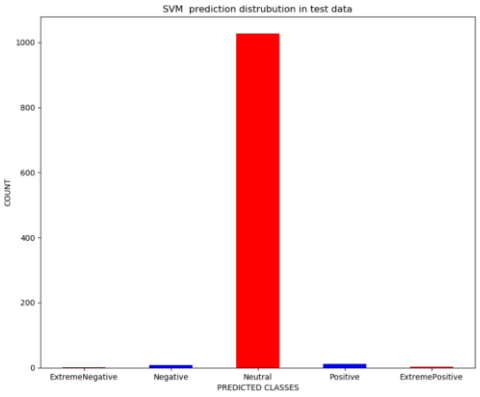


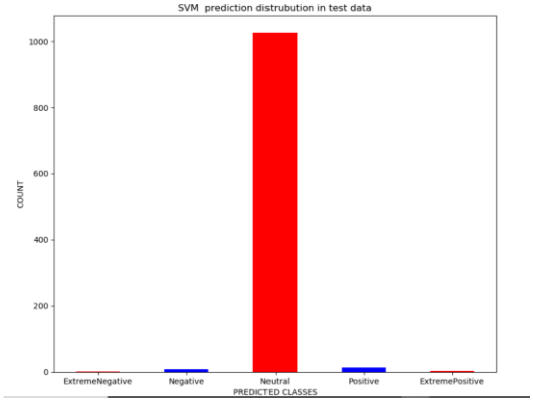


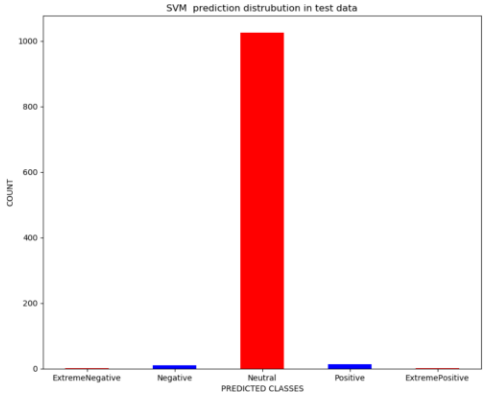


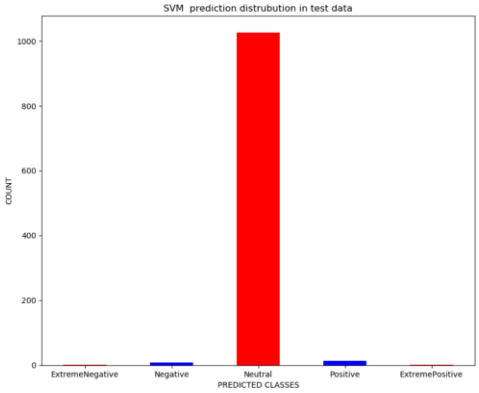
Accuracy graph of SVM for each 10 folds:

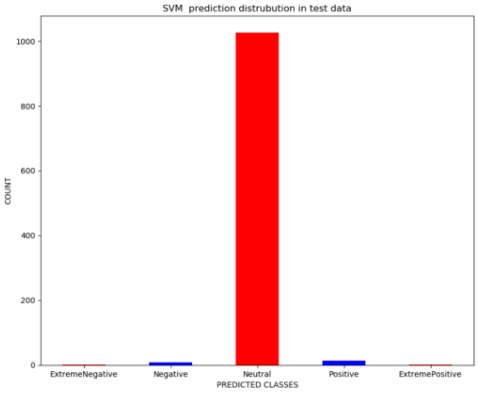
The below graphs show the accuracies of the 10 folds for Naïve Bayes and the values in the results section during validation are taken from these graphs

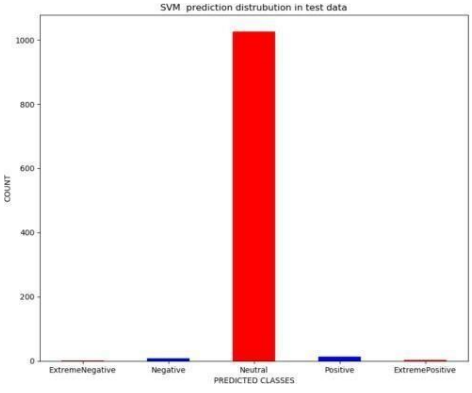

SVM prediction distrubution in test data
COUNT
PREDICTED CLASSES
ExtremeNegative
Negative
Neutral
Positive
ExtremePositive

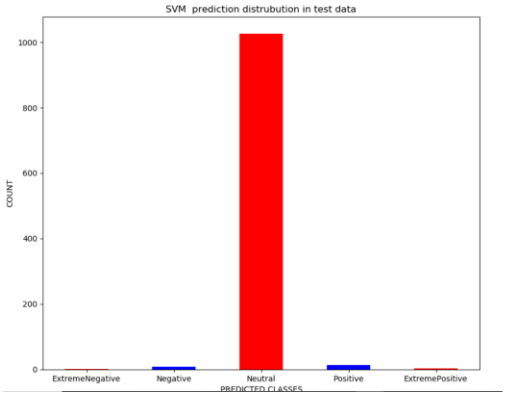

SVM prediction distrubution in test data
COUNT
PREDICTED CLASSES
ExtremeNegative
Negative
Neutral
Positive
ExtremePositive

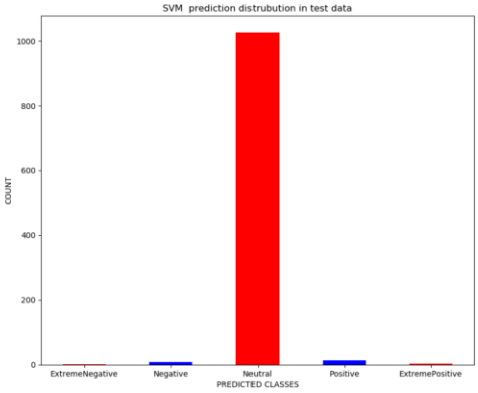

SVM prediction distrubution in test data
COUNT
PREDICTED CLASSES
ExtremeNegative
Negative
Neutral
Positive
ExtremePositive

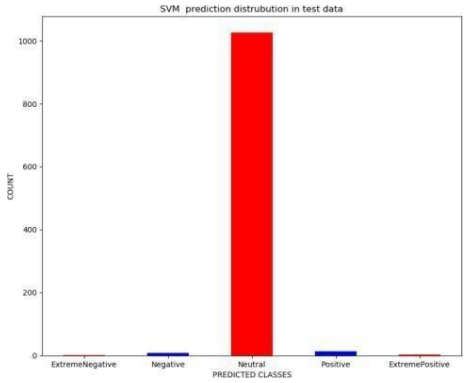

SVM prediction distrubution in test data
COUNT
PREDICTED CLASSES
ExtremeNegative
Negative
Neutral
Positive
ExtremePositive

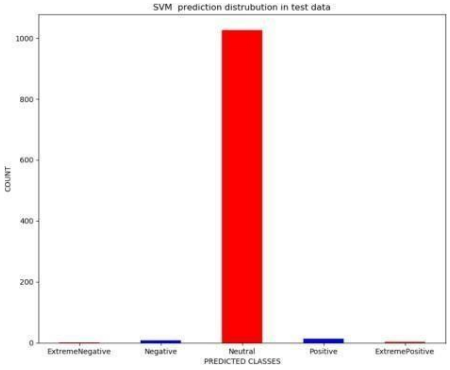

SVM prediction distrubution in test data
COUNT
PREDICTED CLASSES
ExtremeNegative
Negative
Neutral
Positive
ExtremePositive


SVM prediction distrubution in test data
COUNT
PREDICTED CLASSES
ExtremeNegative
Negative
Neutral
Positive
ExtremePositive


SVM prediction distrubution in test data
COUNT
0
200
400
600
800
1000
ExtremeNegative
Negative
Neutral
Positive
ExtremePositive
PREDICTED CLASSES


SVM prediction distrubution in test data
COUNT
0
200
400
600
800
1000
ExtremeNegative
Negative
Neutral
Positive
ExtremePositive
PREDICTED CLASSES


SVM prediction distrubution in test data
COUNT
0
200
400
600
800
1000
ExtremeNegative
Negative
Neutral
Positive
ExtremePositive
PREDICTED CLASSES


SVM prediction distrubution in test data
COUNT
0
200
400
600
800
1000
ExtremeNegative
Negative
Neutral
Positive
ExtremePositive
PREDICTED CLASSES