FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 5 No. 1, March, 2021, pp 566 - 576
DOI: https://doi.org/10.33003/fjs-2021-0501-690

SENTIMENT ANALYSIS OF COVID-19 TWEETS

***Nureni A. Azeez, Ogunlusi E. Victor and Uloko Emmanuel Junior**

Department of Computer Sciences, University of Lagos, Nigeria.

*Corresponding author's Email: nurayhn1@gmail.com

ABSTRACT

Sentiment analysis involves techniques used in analyzing texts in order to identify the sentiment and emotion dominant in such texts and classify them accordingly. Techniques involved include but not limited to preprocessing of texts and the use a machine learning or lexical based approach in classifying these texts. In this research, attempt was made to adopt a machine learning approach to classify tweets on Covid-19 which is considered a global pandemic. To achieve this noble objective, a cross-dataset approach was applied to train four machine learning classification algorithms: Support Vector Machine (SVM), Random Forest (RF) and Naïve Bayes (NB), as well as K-Nearest Neighbors algorithm (KNN). The final result will not only assist us in knowing the best performing algorithm, it will also assist in creating awareness on Covid-19 with the final objective of destigmatizing the patients through the analysis of sentiments and emotions on Covid-19 and finally use the same result for containing the spread of the pandemic.

Keywords: Sentiment analysis, Covid-19, Tweet, Algorithms, Dataset, machine learning

INTRODUCTION

Understanding the sentiment and tone conveyed in a text is really vital especially in the business world as well as in government for decision making. Some areas where this concept is expedient include reviews, ratings and recommendation systems as well as other areas in business. Many businesses have products that have been reviewed by numerous customers and it is paramount to classify these reviews for the purpose of easy accessibility as well as product promotion. Manual labelling and classification of these reviews would be time consuming and inefficient. Sentiment analysis helps to optimize this classification process by automating the entire process.

In the business world, sentiment analysis is not just useful in determining the sentiments of product as it can be used to compare sentiment of competitor's products as well as understanding customer trends based on sentiment values received all year round.

Sentiment analysis also known as opinion mining that finds applications in other areas such as research; as classified information gives a deeper understanding into a study than raw data. Many topics other than product review such as medicine, stock markets, disasters, political topics like elections, social topics like cyberbullying and rape, and many other topics extend the utilization of sentiment analysis (Mäntylä, Graziotin, & Kuutilaa, 2018).

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. It was first recorded in Wuhan, Hubei Province of China in December 2019 (W.H.O., 2020). Currently it has spread to up to 213 countries in six continents (Worldometer, 2020). Coronaviruses are a large family of zoonotic viruses that cause illnesses ranging from common cold to respiratory diseases (Lab-Manager, 2020).

This research attempts to analyze texts sentimentally on the corona virus topic using twitter as a case study. The expected result shall be used for appraisal purposes. It will also be used as the basis for identifying sentiments and emotions of Covid-19 Tweets. It will be used for assessing the impression of individual which will also serve as a platform for sensitizing and creating awareness to the populace on what the pandemic is and what is not.

In summary, this research aims at working on the Sentiment Analysis of COVID-19 Tweets. The objectives to realise from executing the project's aim are as follows:

to ensure a reliable source for the analysis, evaluations, attitude of both positive and negative impression of each aspect contained in the tweets through opinion summarization systems, to guarantee the possibility of providing a distributed and common understanding of COVID-19 pandemic that can be communicated between people of a community and the nation in general, to track, monitor and understand Tweets on COVID-19 in order to have better understanding of the audience. What is more? It is also aimed at keeping on top of what's being said about the pandemic, and discover new trends in the medical and research domains in order to provide valuable insights and thus help governments to formulate effective COVID-19 strategic plans for handling it. It will also assist government to monitor and handle people's grievances. Finally, it will also assist to get in-depth information for strategic analysis and to determine which of the algorithms adopted could provide the best in terms of accuracy and other criteria adopted for evaluation

LITERATURE REVIEW

What is Covid-19? Coronavirus disease (COVID-19) is an infectious disease caused by a novel coronavirus.The first cases of what would later be known as COVID 19 (Corona Virus Identified in December 2019) were recorded in Wuhan China (Huang et al, 2020). The effects of the disease were initially isolated to Wuhan and some places in South East Asia.

The virus was confirmed as Haman-to-human transmissible on 20 January, 2020 after two medical staff were indicted in Guangdong, China. The United States and South Korea also reported their first cases. Things will not remain muted for long, as Italy became the first European country to report cases in January 30, 2020 (WHO, 2020). Germany, Finland, and Italy also reported cases in January. As at 10 May, 2020 there have

been 4 million reported cases and two hundred thousand deaths globally (Huang et al, 2020).

The pandemic has undoubtedly been a topic of conversation across all strata of society. Dominating headlines in all media forms while social media fora have not been different. Twitter for instance has had conversations related to COVID-19 dominating the trends, with opinions supporting or opposing measures taken by different governments to tackle the pandemic. A popular tactic amongst governments across the globe has been to lockdown their economies, to foster social and physical distancing in a bid to curb the spread of the virus. Without work commitments, or the license to work from home, people have taken to the internet to express a myriad of opinions, frustrations and emotions concerning the situation. Deciphering the opinions and general public mood can be very helpful in changing policy direction, changing the way governments relate information concerning the pandemic as they are better positioned to know how some means of information dissipation will be received. The sheer scale of the COVID 19 pandemic and its resultant effect on economies and lives, particularly lockdowns and partial closures of societies has seen a lot of energy expended towards analysing behaviour and reactions (W.H.O., 2020). The following research works are well-related to the subject under discussion.

Neppalli et al. (2017) use Naive Bayes and Support Vector Machine as the two supervised machine-learning classifiers and use a combination of bag-of-words and sentiment features as input to the model.

Gandhe et al. (2018) propose a hybrid approach that combines supervised and unsupervised learning in the sentence-level sentiment analysis model.

Flores (2017) applies sentiment analysis on over 250,000 Tweets to examine the effect of Arizona's 2010 anti-immigrant law on public attitudes towards immigrants. Comparing with traditional survey methods like distributing online polls, an investigation of public sentiments using Twitter data could allow researchers to develop more dynamic responses based on large-scale real-time data, which is tremendously helpful during emergencies.

Dubey (2020) collected Tweets related to coronavirus from March 11 to March 31, 2020 from over ten countries. He suggested that people in France, Switzerland, Netherland, and USA expressed greater distrust and anger compared to other countries such as Italy, Spain, and Belgium. Sentiment analysis can be further combined with topic modeling for a more detailed analysis.

Xue et al. (2020) firstly used the National Research Council of Canada Word-Emotion Association Lexicon, which is a list of English words and their associations with emotions to assign Tweet sentiment by counting the number of words belonging to each emotion category. Then, they applied the Latent Dirichlet Allocation to understand the popular bigrams and sentiments of Tweets.

Samuel et. al in 2020 investigated 9000 tweets and got non-textual variables using N-Gram and further analyzed sentiments using NB, Linear regression, LR and KNN.

Gencoglu in 2020 investigated 26 million tweets using language agnostic BERT sentence embedding models and further classified sentiments using KNN, LR and Bayesian hyperparameter optimization.

Al-rakmi et al. gathered 4,00,000 tweets and implemented entropy and correlation based feature selection and ensemble methods using NB, Bayes Net, KNN, C4.5, random forest (RF) and SVM.

Medford et al, in 2020 studied the early changes in Twitter content, activity and sentiments about COVID 19. This study targeted Twitter users and content made between January 14th to 28th 2020. The study utilised twitter content related to infection prevention practices, vaccination, and racial prejudice. Measurements were made while experimentation to decipher opinions on emotional violence and predominant emotions were carried out. The study discovered that the hourly mentions of COVID19-related keywords increased exponentially from 21st January, 2020. Half of the studied content expressed fear and thirty percent expressed surprise. Racially charged tweets mirrored the diagnosis of new cases of COVID-19. Of the 126, 049 evaluated tweets, the economic and political impact of the pandemic was the most common topic of conversation, while the prevention and health risks to the public associated with the spread of the disease were least discussed (Medford et al, 2020).

The impact of the pandemic on the mental health of the citizens and also on the economies of countries worldwide that had taken the approach of a lockdown to curb the spread of COVID 19 is another area of concern. Barkur et al, carried out a sentiment analysis study of Indians after lockdown announcements were made public. Twitter was the platform from which data was sourced. Tweets, streamlined to two popular hashtags (#IndiaLockdown and #IndiafightsCorona) from the period of March 25th to 28th were studied to gauge the feelings of Indians towards the lockdown. A word cloud aggregating the prevailing emotions and opinions conveyed in the tweets was generated. The study found that despite the gloom about what the pandemic may spell for the 1.3 billion dense population of India, Indians were positive, showing optimism about flattening the curve (Barkur et al, 2020).

Li et al, in 2020 studied the difference in psychological profile of Weibo posters in Easter China. Before and after January 20th (classified as a type B infectious disease by the Chinese National Health Commission) (Li et al, 2020). The samples used in the study were from the Weibo user pool, containing 1.16 million active users. User profile information, network information and messages were included as part of the study.

Around eighteen thousand users were selected and their original posts fetched. Online Ecological Recognition which provides a means of automatically deducing psychological profile by using machine learning predictive modelling was employed on the data. TextMind system was used to extract content features. The study found an increase in topics concerning health and family with a decrease in topics concerning leisure and friends. Posters showed more negative emotions (anxiety, depression, and indignation) and less positive emotions after the declaration (Li et al, 2020).

In 2020, Dubey identified the prevailing sentiments of citizens of twelve countries (USA, Italy, Spain, Germany, China, France, UK, Switzerland, Belgium, Netherlands, Australia and India) regarding COVID 19 from the period from March 11th to 31st 2020. After scoring the tweets on the basis of sentiments and emotions, the word cloud for each country was developed. The study showed varying degree of positivity concerning COVID 19 amongst countries with Belgium and India being the most positive, while the United States, Switzerland and China showed the most negative emotions (Dubey 2020).

Going further to analyse emotions laced in collected tweets, USA, France and China were found to have tweeted the most

with anger, Switzerland tweeted the most with sadness and fear, while trust and surprise were mostly exhibited by Belgium.
In this work, efforts were made to apply a cross-dataset approach to train four machine learning classification algorithms: Support Vector Machine (SVM), Random Forest (RF) and Naïve Bayes (NB), as well as K-Nearest Neighbors algorithm for this purpose. The results obtained were used to make comments and draw the final conclusion. Table 1 provides the results as well as approaches adopted by similar works in a couple of literatures reviewed by the authors.

Table 1: Summary of results and methods used by related works

Title	Author(s)	Method (s)	Results
Sentiment analysis during hurricane Sandy in emergency response	Neppalli et al. (2017)	Naive Bayes and Support Vector Machine	Extraction of sentiments during a disaster may help emergency responders develop stronger situational awareness of the disaster zone itself.
Sentiment analysis of twitter data with hybrid learning for recommender applications	Gandhe et al. (2018)	supervised and unsupervised learning	The results of this work would be helpful in providing recommendations to users for product reviews, political campaigns, stock predictions, urban policy decisions.
Twitter sentiment analysis during COVID19 outbreak	Dubey (2020)	Topic modeling	The results of the study concludes that while majority of the people throughout the world are taking a positive and hopeful approach, there are instances of fear, sadness and disgust exhibited worldwide.
Twitter discussions and emotions about COVID-19 pandemic: A machine learning approach.	Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Liu, X., & Zhu, T (2020)	Latent Dirichlet Allocation	This study showed that Twitter data and machine learning approaches can be leveraged for an infodemiology study, enabling research into evolving public discussions and sentiments during the COVID-19 pandemic.
COVID-19 public sentiment insights and machine learning for tweets classification, Information	Samuel et. al in 2020	NB, Linear regression, LR and KNN	This research provides insights into Coronavirus fear sentiment progression, and outlines associated methods, implications, limitations and opportunities.
Large-scale, language-agnostic discourse classification of tweets during COVID-19, Mach. Learn. Knowl. Extraction	Gencoglu, O. (2020)	KNN, LR and Bayesian hyperparameter optimization	large-scale surveillance of public discourse is feasible with computationally lightweight classifiers by out-of-the-box utilization of these representations
Lies kill facts save: Detecting COVID-19 misinformation in Twitter	Al-Rakhami, M.S and Al-Amri, A.M.(2020)	NB, Bayes Net, KNN, C4.5, random forest (RF) and SVM.	The results obtained with the proposed framework reveal high accuracy in detecting credible and non-credible tweets containing COVID-19 information.
An "Infodemic": Leveraging High-Volume Twitter Data to Understand Public Sentiment for the COVID-19 Outbreak. (2020)	Richard J. Medford, Sameh N. Saleh, Andrew Su marsono, Trish M. Perl, Christoph U. Lehmann	Recurrent neural networks to label emotion for a document according to Ekman's emotional classification	Tweets with negative sentiment and emotion parallel the incidence of cases for the COVID-19 outbreak.
Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. Asian Journal of Psychiatry. (April 2020)	Barkur G, Vibha, Kamath GB	Data and content analysis	Indians overall positive about flattening the curve, some concern about livelihoods of day laborers
The Impact of COVID-19 Epidemic Declaration on Psychological Consequences: A Study on Active Weibo Users (2020)	Li, S.; Wang, Y.; Xue, J.; Zhao, N.; Zhu, T.	Online Ecological Recognition for psychological profile evaluation and Text Mind for feature extraction.	Posters on Weibo showed more negative sentiments after the declaration that COVID 19 was a type B infectious disease and figures about its mortality rate were announced
Twitter Sentiment Analysis during COVID-19 Outbreak (April 9, 2020)	Dubey, A. D.,	Data and content analysis	There are instances of fear, sadness and disgust exhibited worldwide. France, Switzerland, Netherland and United States of America have shown signs of distrust and anger, compared to other countries that have shown positive, hopefulness.

As at 29th, May 2020, statistics shows that the active covid-19 cases has a total number of 2, 915, 882 as those in Mild Condition which represents 98% of the infected patients while 52, 947 which is 2% signifies those in serious or critical condition (Worldometer, Covid-19 Coronavirus Pandemic, 2020). Conversely, for the closed cases, a total of 2, 593, 676 which represents 88% patients have recovered and discharged while 362, 555 which is 12% represents the total death rate as at 29th, May 2020.

METHODOLOGY OF THE PROPOSED METHOD OF SENTIMENT ANALYSIS

Figure 1 presents the basic diagrammatic representation of the system architecture utilized for the sentiment analysis of covid-19 tweets.

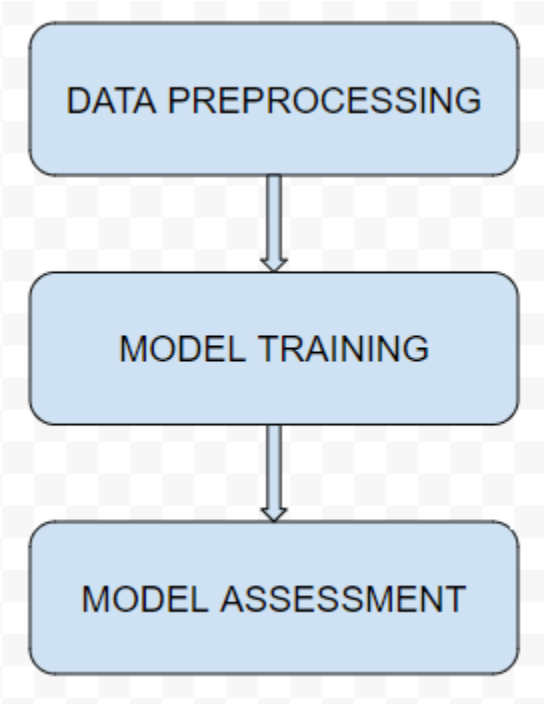

Fig. 1. System Architecture of the proposed Covid-19 sentiment analyzer.

MACHINE LEARNING ALGORITHMS USED

In this research work, four different algorithms (classifiers) were used. They are discussed as follows (Azeez et. al., 2020):

1. **Support Vector Machine (SVM) -** It is a supervised learning model with associated learning algorithms that gives analysis of data for classification. It represents the data as points in space. The classification into individual groups is achieved by discovering the best hyperplane that distinguishes the two classes in the optimal approach. Support Vector Machine separates positively labeled examples from the negatively labeled ones by finding the "hyperplane $w^T x = 0$" that maximizes margin between the two classes which can be achieved by solving quadratic objective function (Azeez et. al., 2021):

$$h : w^T x + b = 0 \quad \ldots\ldots(1)$$

Where *b* is the intercept and bias term of hyperplane equation, $w^T x = 0$ is the hyper plane.

2. **Random Forest** – Random Forest algorithm does the selection of observation and feature randomly in order to build several decision trees and then computes the average of the results. Random Forest algorithm creates random subset of the features and builds smaller trees using the subset created. Furthermore, Random Forest produces high accuracy through cross validation, handles missing values and maintains the accuracy of large proportion of data. Random Forest classifiers don't allow over-fitting trees into the model in case there are no more trees (Azeez et. al., 2021).

$$RFfi_i = \frac{\sum_j norm\, fi_{ij}}{\sum_{j \in all\ features, k \in all\ trees} normfi_{jk}} \quad \ldots\ldots\ldots\ldots (2)$$

Where RF f*i* sub (*i*) is the importance of feature *i* calculated from all trees in the random forest model. F*i* sub (*i*, *j*) are the importance of feature for node of *i* and *j*

3. **Naïve Bayes Classifier:** The purpose of using a Naïve Bayes Classifier is to predict the likelihood that an event will occur with the assistance of evidence that is present in the data. A multinomial Naïve Bayes algorithm classifier was used because it is suitable and more efficient for features that describe discrete frequency counts which is similar to the features of the data present in the dataset obtained (Azeez et. al., 2021).

Given a class variable or hypothesis (y) and a dependent feature or evidence $(x1, \dots, xn)$
Therefore,

$$P(y|x1, x2, x3 \dots . xn) = \frac{P(y)P(x1,x,x3,\dots\dots xn|y)}{P(x1,x,x3,\dots\dots xn)} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

where: $P(y)$ *are labels*
$P(x)$ *are comments*

$P(y|x1, x2, x3 \dots . xn)$ is how probable was the hypothesis (labels) given the observed evidence (Zhang, H. 2004).

$P(x1, x2, x3 \dots . xn|y)$ is how probable is the evidence, given that the hypothesis is true.

$P(y)$ is how probable was the hypothesis before observing the evidence.

$P(x1, x2, x3 \dots \dots xn)$ is how probable is the new evidence under all possible hypothesis.

4. K-Nearest Neighbor (**KNN) –** is a parametric method that is used for classification. An object is classified by plurality vote of its neighbor with the object being assigned to the class most common among its K nearest neighbors. A commonly used distance metric for continuous variables is Euclidean distance (Azeez et. al., 2019).

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \dots\dots\dots\dots\dots\dots (4)$$

$$= \sqrt{\sum_{i=1}^{n}(qi - Pi)^2} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (5)$$

where q_1 *to* q_n *represents the attributes value for one observation*
p_1 *to* p_n *represents the attribute value for the other observation*

DATA DESCRIPTION AND PREPROCESSING

Sentiment140 dataset which contains up to 1.6 million tweets was selected for training the models. It's a dataset that was automatically classified by taking advantage of the emojis present in tweets. (Go et. al., 2009). The sentiment140 dataset can be gotten through the link https://www.kaggle.com/kazanova/sentiment140 on Kaggle.

The columns of concern in the dataset are the tweet and the target. The target was modified to use a 0-1 range with '0' indicating a negative text and '1' indicating a positive text, while the text column was preprocessed. Preprocessing on the data involved converting the texts to lowercase, removal of punctuation and vectorizing the texts.

The process of vectorizing the text involves getting all unique words contained in all the tweets in the train data, filtering those words and then creating a feature index for each remaining unique word. With this feature indices, instances in the train data are then defined, with each feature index containing the term frequency-inverse document frequency (TFIDF) score for that word in the instance. The TFIDF score serves as a weight for each word and signifies the importance of each word in an instance.

The TFIDF score is calculated as follows:

$$\boldsymbol{tf}(\boldsymbol{t},\boldsymbol{d}) = \frac{count\ of\ t\ in\ d}{number\ of\ words\ in\ d} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (6)$$

$$\boldsymbol{df}(\boldsymbol{t}) = count\ of\ t\ in\ all\ documents \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (7)$$

$$\boldsymbol{idf}(\boldsymbol{t}) = \ln\left(\frac{N+1}{df+1}\right) + 1 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (8)$$

$$\boldsymbol{tfidf}(\boldsymbol{t},\boldsymbol{d}) = tf \times idf \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (9)$$

The TFIDF value is then normalized using l2 normalization technique which normalizes all features in an instance using the square root of the sum of squared vectors in that instance. It is described mathematically as follows:

$$\boldsymbol{tfidf}_{\boldsymbol{normalized}} = \frac{tfidf(t,d)}{\sqrt{\sum_{i=1}^{n} tfidf_i{}^2}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (10)$$

Where

t *is a term (word)*
d *is a document (instance)*
tf *is the term frequency of* ***t*** *in a document* ***d***
N *is the number of documents*

df** is the document frequency of a term **t
idf** is the inverse document frequency of a term **t
tfidf** is the term frequency-inverse document frequency of a term **t** in a document **d
*$tfidf_{max}$ is the maximum **tfidf** score of that term*
***n** is the number of terms/features*
*$tfidf_i$ is the ith vector in an instance/document **d***

Stop words were not explicitly removed from the vectorization process, as stop words sometimes are important in the context of a statement. However, we relied on the following rules to remove noisy terms from our data.

a. Terms that appear less than 5 terms in all documents (i.e. df less than 5).
b. Terms that appear more than more than 75 times in every 100 documents (i.e. **df** /number of documents greater than 0.75). This rule implicitly removes certain stop words.
c. Terms not starting with an English alphabet e.g. '2aa'.
d. Terms with less than 3 characters.

By following the above procedure, we achieved data vectorization on 2618 terms.

1. MODEL TRAINING

The preprocessed data was used to train four machine learning models namely: Support Vector Machine (SVM), Random forest (RF) and Naïve Bayes (NB), and also K-Nearest Neighbors algorithm. The sentiment 140 sample dataset was split in the ratio 0.85 to 0.15 with the former used to train the models and the latter used to test them.

2. MODEL ASSESSMENT

After preprocessing the data, and training models, assessment of the trained models was in three stages:

a. Sentiment140 dataset test data.
b. IMDB movie review dataset test data. (Maas, et al., 2011). It is a dataset containing various movie reviews and their sentiments on the International Movie Database (IMDB). It can be downloaded from Kaggle with the link https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.
c. Covid-19 dataset test data. Raw unlabeled covid-19 twitter dataset can be gotten on Kaggle with the link https://www.kaggle.com/smid80/coronavirus-covid19-tweets.

3. MODEL PERFORMANCE MEASURES

Cross dataset testing with the IMDB movie review dataset was performed iteratively together with model training in order to assess model performance and make modifications to the models. Measure of model performances were based on the following statistics:

True Positive (TP) – positive instances that are correctly predicted as positive. It is given as follows:

$$\boldsymbol{TP}_{\%} = \frac{correctly\ predicted\ p}{total\ number\ of\ p} \times 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (11)$$

Where
***p** = positive instances*

False Negative (FN) – positive instances that are incorrectly predicted as negative. It is given as:

$$\boldsymbol{FN}_{\%} = \frac{incorrectly\ predicted\ p}{total\ number\ of\ p} \times 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (12)$$

Or

$$\boldsymbol{FN}_{\%} = 100 - TP \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (13)$$

Where
***p** = positive instances*
***TP** = True Positive*

True Negative (TN) – negative instances that are correctly predicted as negative. It is given as:

$$\boldsymbol{TN}_{\%} = \frac{correctly\ predicted\ n}{total\ number\ of\ n} \times 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (14)$$

Where
***n** = negative instances*

False Positive (FP) – negative instances that are incorrectly predicted as positive. It is given as:

$$\boldsymbol{FP}_{\%} = \frac{incorrectly\ predicted\ n}{total\ number\ of\ n} \times 100 \ldots \ldots \ldots \ldots \ldots \ldots (15)$$

Or

$$\boldsymbol{FP}_{\%} = 100 - TN \ldots \ldots \ldots \ldots \ldots \ldots (16)$$

Where
***p** = negative instances*
***TN** = True Negative*

Accuracy – Basic measure of model correctness on predictions.

$$\boldsymbol{A}_{\%} = \frac{correctly\ predicted\ i}{total\ number\ of\ i} \times 10 \ldots \ldots \ldots \ldots \ldots \ldots (17)$$

Or

$$\boldsymbol{A}_{\%} = \frac{TP_{\%} + TN_{\%}}{TP_{\%} + FN_{\%} + TN_{\%} + FP_{\%}} \ldots \ldots \ldots \ldots \ldots \ldots (18)$$

Where
***i** = instances*
***TP** = True Positive*
***TN** = True Negative*
***FP** = False Positive*
***FN** = False Negative*

Precision – defined as the fraction of relevant instances among the retrieved instances. It is calculated as follows:

$$\boldsymbol{P_P} = \frac{TP}{TP + FP} \ldots \ldots \ldots \ldots \ldots \ldots (19)$$

$$\boldsymbol{P_N} = \frac{TN}{TN + FN} \ldots \ldots \ldots \ldots \ldots \ldots (20)$$

Where
***TP** = True Positive*
***TN** = True Negative*
***FP** = False Positive*
***FN** = False Negative*

Recall – defined as the total amount of relevant instances that were actually retrieved. It is given as:

$$\boldsymbol{R_P} = \frac{TP}{TP + FN} \ldots \ldots \ldots \ldots \ldots \ldots (21)$$

$$\boldsymbol{R_N} = \frac{TN}{TN + FP} \ldots \ldots \ldots \ldots \ldots \ldots (22)$$

Where
***TP** = True Positive*
***TN** = True Negative*
***FP** = False Positive*
***FN** = False Negative*

F_1 score – defined as the harmonic mean of precision and recall. It is calculated as follows:

$$\boldsymbol{F_1} = \boldsymbol{2} \times \frac{P \times R}{P + R} \ldots \ldots \ldots \ldots \ldots \ldots (23)$$

Where
***P** = Precision*

R = *Recall*

THE RESULTS OBTAINED

The results obtained are summarized in Table 2. The graphical representations for each of the Sentiment, IMDB data and Covid-19 Tweets are represented in Figures 2, 3 and 4 respectively.

Table 2. Summary of results obtained from SVM model on Sentiment 140 data

Metrics	Value Obtained for Datasets		
	Sentiment	**IMDB data**	**COVID-19 TWEETS**
Accuracy (%)	77.72	71.46	77.56
True Positive (%)	75.63	58.06	81.82
True Negative (%)	79.53	84.0	74.07
False Positive (%)	20.47	16.0	25.93
False Negative (%)	24.37	41.94	18.18
Precision	0.78	0.73	0.78
Recall	0.78	0.71	0.78
F1-score	0.78	0.71	0.78

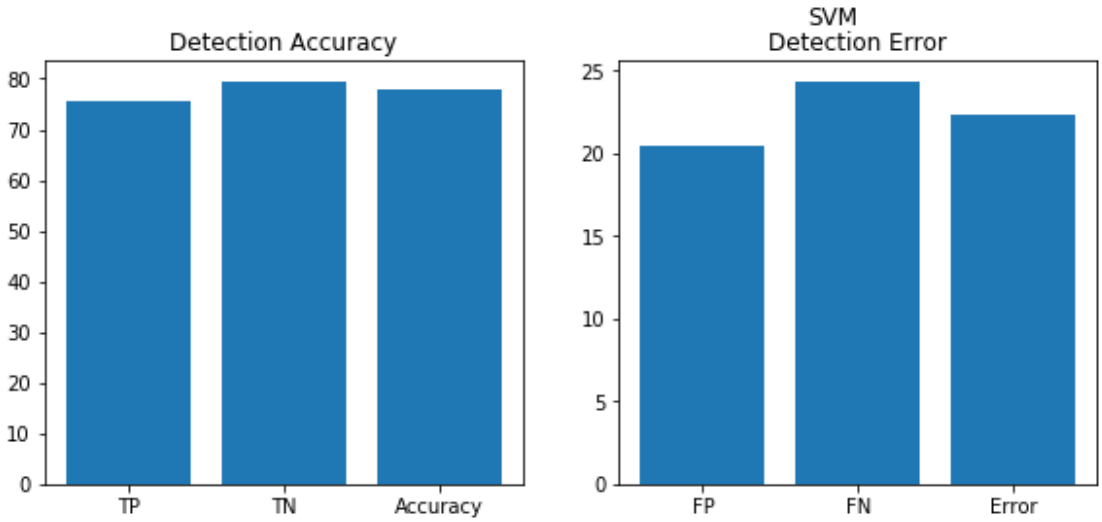

Fig. 2. Detection accuracy and detection error graph on sentiment140 test data with SVM.

TEST WITH SENTIMENT140 TEST DATA

Testing the SVM model with sentiment140 test data yielded an accuracy of 77.72%, a true positive of 75.63% and a true negative of 75.93%. The precision, recall and F1-score all settle in at 0.78. The False Negative value of 24.37% is the highest error value as indicated in the figure above.

TEST WITH IMDB MOVIE REVIEW DATA (SVM)

Results on IMDB movie review data yielded a True Negative of 84%. However, the True Positive value was 58%. This could be due to the fact that each movie review text is long and usually contains a critical element even for positive reviews. The results are summarized in Table 2.

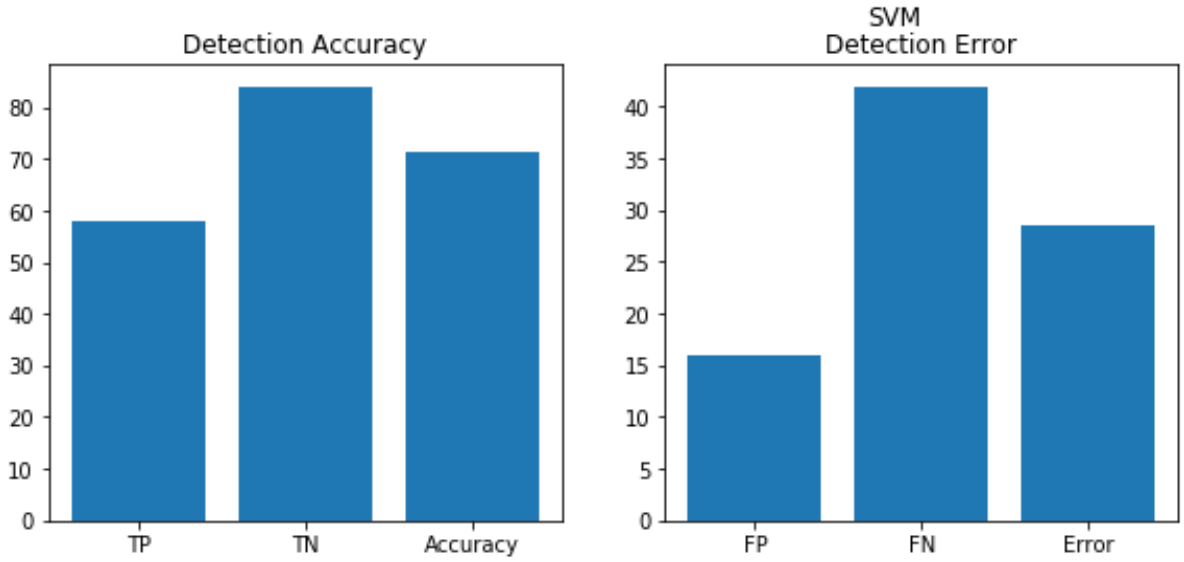

Fig. 3. Detection accuracy and detection error graph on IMDB movie review test data with SVM.

With the IMDB test data, there is higher accuracy on the negative side than the positive with true negative value yielding 84% accuracy. The highest error value was the False Negative with a high value of 41.94% as a result of many positive reviews being classified as negative.

TEST WITH COVID-19 TWEETS USING SVM

The sample dataset consists of messages in English language tweeted on the 27th of May, 2020 and having one of the following hashtags: #coronavirus, #coronavirusoutbreak, #coronavirusPandemic, #covid19, #covid_19. The results obtained on sample covid-19 tweets are summarized in Table 2

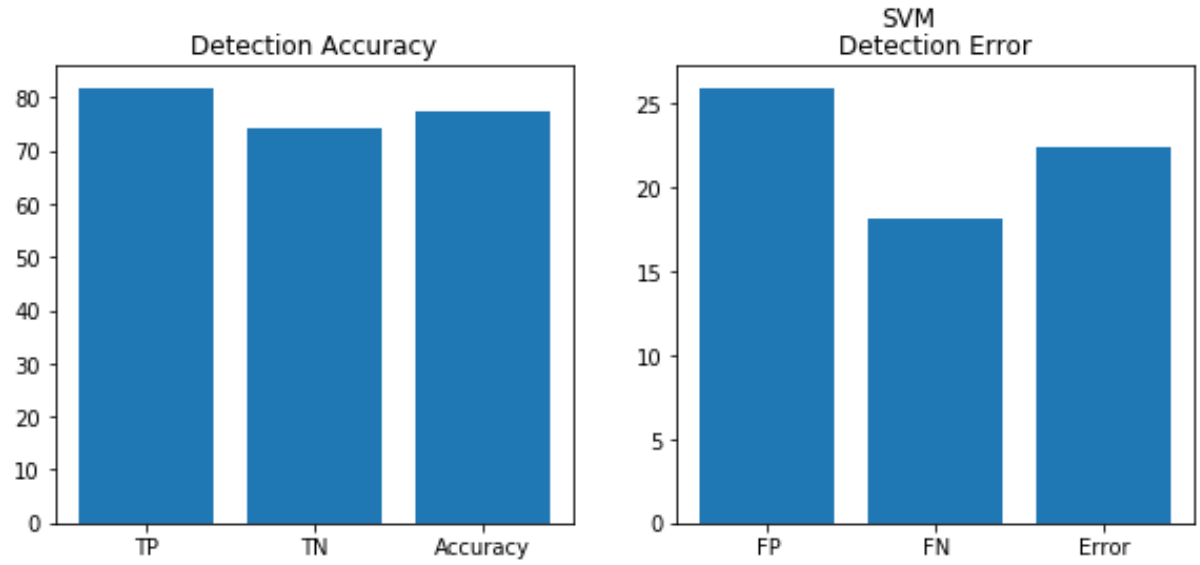

Fig. 4 Detection accuracy and detection error graph on sample covid-19 test data with SVM.

Results favor positive tweets as a true positive value of 81.82%, a true negative value of 74.07% and an accuracy 77.56% was achieved. The highest detection value was the False positive with a value of 25.93%.

COMPARISON OF SVM WITH OTHER MODELS

Table 3. Summary of results obtained from SVM on all models

		SVM			RF			NB			KNN	
	SEN	IMDB	COV	SEN	IMDB	COV	SEN	IMDB	COV	SEN	IMDB	COV
Accuracy (%)	77.72	77.72	77.55	73.64	57.9	69.39	71.08	60.22	65.31	72.47	61.62	55.1
True Positive (%)	75.63	58.06	81.82	70.02	52.03	77.27	77.19	81.6	63.64	70.73	61.54	59.09
True Negative (%)	79.53	84.0	74.07	76.77	63.4	62.96	65.02	40.02	66.67	74.19	61.7	51.85
False Positive (%)	20.47	16.0	25.93	23.23	36.6	37.04	34.98	59.8	33.33	25.81	38.3	48.15
False Negative (%)	24.37	41.94	18.18	29.98	47.97	22.72	22.81	18.4	36.36	29.27	38.46	40.91
Precision	0.78	0.73	0.78	0.74	0.58	0.71	0.71	0.63	0.65	0.72	0.62	0.52
Recall	0.78	0.71	0.78	0.74	0.58	0.69	0.71	0.60	0.65	0.72	0.62	0.55
F1-score	0.78	0.71	0.78	0.74	0.58	0.69	0.71	0.59	0.65	0.72	0.62	0.55

The SVM model has the highest accuracy on all test datasets with an average accuracy of 77.66%. The Random Forest model comes in second with an average accuracy of 66.98%, Naïve Bayes model third with 65.54% and K-Nearest Neighbors algorithm finally with an average of 63.06%. The values obtained for each of the models are presented in Table 3.

CONCLUSION

In this research, efforts were made to train four classification models: Support Vector Machine, Random Forest, Naïve Bayes and K-Nearest Neighbors algorithms on the sentiment140 data in order to have a general sentiment analysis model that can classify covid-19 tweets. Cross dataset testing with IMDB movie review dataset was performed and tested the models with sample covid-19 tweets. With the Support Vector Machine model, a True Positive of 81.82% was obtained, a True Negative of 74.07% and an accuracy of 77.55% on the sample covid-19 dataset.

One of the limitations of this work is that interrogative sentences, questions and sarcastic statements are very likely to yield undesirable results. Also based on training data used, it is a binary classification with no provisions for neutral statements. The classification is limited only to the English language and is very likely to give wrong results for other languages. Another issue is the cross-dataset training approach, as models specifically trained on the covid-19 topic may have better insight in the classification of certain tweets although with certain trade-offs. An investigation into such approach in the future is foreseeable.

REFERENCES

Go A., Bhayani R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *Processing*, 150.

Lab-Manager. (2020, 03 16). *COVID-19: A History of Coronavirus*. Retrieved from Lab Manager: https://www.labmanager.com/lab-health-and-safety/covid-19-a-history-of-coronavirus-22021

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 142-150). Portland: Association for Computational Linguistics.

Mäntylä, M. V., Graziotin, D., & Kuutilaa, M. (2018, February). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review, 27*, 16-32.

W.H.O. (2020, 04 17). *Q&A on coronaviruses (COVID-19)*. Retrieved from World Health Organization: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses

Worldometer. (2020, 05 27). *Countries where COVID-19 has spread*. Retrieved from worldometer: https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/

Worldometer. (2020, 05 28). *Covid-19 Coronavirus Pandemic*. Retrieved from Worldometer: https://www.worldometers.info/coronavirus/

Medford, R.J., Saleh, N.S., Sumarsono A, Perl M. T., Lehmann U. C. (2020). An "Infodemic": Leveraging High-Volume Twitter Data to Understand Public Sentiment for the COVID-19 Outbreak

Barkur G, Vibha, Kamath GB. Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. Asian Journal of Psychiatry. 2020 Apr; 51:102089. DOI: 10.1016/j.ajp.2020.102089.

Li, S.; Wang, Y.; Xue, J.; Zhao, N.; Zhu, T. The Impact of COVID-19 Epidemic Declaration on Psychological Consequences: A Study on Active Weibo Users. Int. J. Environ. Res. Public Health 2020, 17, 2032.

Dubey, A. D., Twitter Sentiment Analysis during COVID-19 Outbreak (April 9, 2020).

Huang, Chaolin, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang. "Clinical features of patients

infected with 2019 novel coronavirus in Wuhan, China." The Lancet 395, no. 10223 (2020): 497-506. 2.

(WHO), World Health Organization. "Preliminary Investigations Conducted by the Chinese Authorities Have Found No Clear Evidence of Human-to-Human Transmission of the Novel #Coronavirus (2019- NCoV) Identified in #Wuhan, #Chinacn. Pic.twitter.com/Fnl5P877VG." Twitter, Twitter, 14 Jan. 2020, twitter.com/WHO/status/1217043229427761152.

NA. Azeez, OE Adio, AW Yekinni and CJ Onyema (2020) " Evaluation of Machine Learning Algorithms for Filtering and Isolating Spammed Messages" FUTA Journal of Research in Sciences, Vol. 16(1), April, 2020: 26-38

N.A. Azeez, S.O. Idiakose, C.J. Onyema, and C.V Vyver (2021) "Cyberbullying Detection in Social Networks: Artificial Intelligence Approach" Journal of Cyber Security and Mobility, Vol. 10 4, 1–30. doi: 10.13052/jcsm2245-1439.1046

N.A Azeez, Ihotu Agbo Margaret, Misra Sanjay (2021) "Adopting Automated White-List (AWL) Approach for Anti-Phishing Solution" Elsevier Journal of Computers & Security 108 (2021) 102328, pp. 1-18

Neppalli, V. K., Caragea, C., Squicciarini, A., Tapia, A., & Stehle, S. (2017). Sentiment analysis during hurricane Sandy in emergency response. International Journal of Disaster Risk Reduction, 21, 213–222.

Gandhe, K., Varde, A. S., & Du, X. (2018). Sentiment analysis of twitter data with hybrid learning for recommender applications. In , 2018. 2018 9th IEEE annual ubiquitous computing, Electronics & Mobile Communication Conference (UEMCON), New York City, NY, USA (pp. 57–63). https://doi.org/10.1109/UEMCON.2018.8796661.

Flores, R. D. (2017). Do anti-immigrant laws shape public sentiment? A study of Arizona's SB 1070 using twitter data. American Journal of Sociology, 123(2), 333–384.

Dubey, A. D. (2020). Twitter sentiment analysis during COVID19 outbreak. Available at: https://ssrn.com/abstract=3572023.

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Liu, X., & Zhu, T (2020). Twitter discussions and emotions about COVID-19 pandemic: A machine learning approach. (2020). arXiv:2005.12830.

Samuel, J., Ali, G.G.M.N., Rahman, M.M., Esawi, E., Samuel, Y. (2020) COVID-19 public sentiment insights and machine learning for tweets classification, Information 11 (2020) 314, http://dx.doi.org/10.3390/info11060314, URL: https://www.mdpi.com/2078-2489/11/6/314, number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

Gencoglu, O. (2020) Large-scale, language-agnostic discourse classification of tweets during COVID-19, Mach. Learn. Knowl. Extraction 2 (2020) 603–616, http://dx.doi.org/10.3390/make2040032, URL: https://www.mdpi. com/2504-4990/2/4/32, number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Al-Rakhami, M.S and Al-Amri, A.M.(2020) Lies kill facts save: Detecting COVID-19 misinformation in Twitter, IEEE Access 8 (2020) 155961–155970, http: //dx.doi.org/10.1109/ACCESS.2020.3019600, conference Name: IEEE Access.

Azeez, N.A, Salaudeen, B.B, Misra, S. Damasevicius, R. Maskeliunas, R. (2019) "Identifying Phishing Attacks in Communication Networks using URL Consistency Features" International Journal of Electronic Security and Digital Forensics (InderScience). https://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijesdf