Analysing Sentiments on Covid19 (Corona Virus) Tweets by Comparing the Accuracies using Word Embeddings & Machine Learning Models

Vedhosi Vaibhav Shashidhara - 19210927, Bharath Reddy Nagasetty - 19211306,
Sachin Mahesh - 19211191, Bidish Basu - 19210427

Department of Computing
Dublin City University
Dublin, Ireland
vedhosi.shashidhara2@mail.dcu.ie, bharath.nagasetty2@mail.dcu.ie,
sachin.mahesh3@mail.dcu.ie, bidish.basu2@mail.dcu.ie
Github Link - https://github.com/vedhosivaibhav/Sentiment-Analysis-Covid19

***Abstract*— Social media nowadays is at a boom in the recent times which it is an internet-based form of communication where large amount of data is generated having millions of users. Considering the fact that people spend hours daily on social media and share their opinion on various different topics helps us analyze sentiments better. Some of the social networking sites such as Instagram, Twitter and Facebook give out information about people's sentiments and opinions by providing analysts a platform to collect raw data. Sentiment analysis of social media data consists of attitudes, assessments, and emotions which can be considered a way human think. In a large collection of documents understanding and classifying the polarities are a very difficult task. This paper aims to perform sentiment analysis on Covid19 Tweets by comparing the accuracies using word embeddings such as Word2Vec and Doc2Vec with Machine Learning Models such as Logistic regression, support vector machine and Random Forest. The traditional methods such as Term-Frequency Inverse Document Frequency (TF-IDF) and the Bag of Words (BoW) will be compared with the word embedding models. The results show that Word2vec with Random Forest for big data improves the accuracy of sentiment analysis by considering contextual semantics of words in the text.**

Keywords—Sentiment analysis; Logistic Regression; Support Vector Machine; Random forest; Word2Vec; Doc2Vec; TF-IDF; BoW

I. INTRODUCTION

Social media plays an important role in day to day routine and the data generated is not a joke in this digital era, where every second there are large amounts of real time data that are generated across the globe that are majorly in the form of unstructured data format. Using these unstructured data generated from the social media, the sentiments and opinions of people can be known. Twitter is one of the biggest contributors of big data where it also provides views on Covid19 by the people. Any news or emerging events are almost instantly followed thereby causing a burst in Twitter volume and providing a unique opportunity to find the essence the relationship between Covid19 and public sentiment. Twitter currently has 330 million monthly active users and thereby is accessible through SMS, mobile devices and website interface. Twitter provides a platform for the users to deliver, interpret and share 280 characters post which is known as a tweet. Moreover, 80% of its current users are active through mobiles. The natural language processing (NLP) plays a major role in twitter data analysis and can be used to analyze unstructured tweets to identify their opinions [1]. The microblogging service users like Twitter users tend to make spelling mistakes by typing the tweet and the fact that they try to use emoticons for expressing their views and emotions along with text.

One of the challenging tasks in machine learning is the numeric representation of a text document. There are a couple of well-known existing methods such as Bag Of Words (BoW) and the Term-Frequency-Inverse Document Frequency (TF-IDF) where the results obtained from these aren't that great, the reason being considering the ordering of the word, BoW loses a lot of the subtleties of a possible good representation.

In Large scale real time sentiment analysis of text feature selection method plays important role in improving the accuracy. Such as Doc2Vec and Word2Vec models, a distributed representation of word is a set of the unsupervised shallow two-layer neural network model that produces word embedding. Word2Vec considers contextual semantics of words to produce word embedding i.e., instead of focusing on single word and two or three word it considers the context in which the word is occurring. Similar words with same or relative context are mathematically clustered together into vector space. That further conserves the semantic relationship between words. Numeric description of word in the form of vector is known as Word embedding. The Doc2vec model, regardless of the length creates a numeric expression of a document. But unlike words, documents do not come in logical structures such as words. Hence word embedding produced using Word2Vec, can be used to train machine learning and classification algorithm to improve sentiment calcification accuracy based on context and semantic relation between words.

Logistic regression is one of the classification algorithms used to assign observations to a discrete set of classes. Logistic regression returns a probability value by transforming its output using the logistic sigmoid function to which it can then be mapped to two or more discrete classes. Support Vector Machine is a supervised machine learning algorithm where the data items are considered as points in a multi-dimensional space. The classification is performed by constructing the optimal hyper-planes, differentiating the classes well by maximizing the margin between the class's closest points. Random Forest is one of a versatile machine learning algorithm, capable of performing both classification and regression tasks. Random Forest is an ensemble learning

model, where a few weak models combine to form a powerful model. In Random Forest, one grows multiple trees as opposed to a single decision tree.

II. Related Work

The social media data has valuable, vast and rapidly emerging unstructured information which has created an opportunity to study public opinion & know the sentiments of the people. Adaption to change quickly is very crucial in this ever-changing world for taking any actions or decisions faster. Faster the data is available, faster an action or decision(s) [1] could be taken and in cases where the action has likely saved the lives and by preventing the dropping of lives.

A. Literature Survey

Efthymios Kouloumpis, Theresa Wilson, Johanna Moore [4] in the year 2011 published a research paper on Twitter Sentiment analysis. In this paper [4] the methods used were N-gram, Lexicon feature and PoS to perform sentiment analysis where it was found that PoS may not be useful for microblogging sentiment analysis. Choosing a supervised approach for solving a problem is always a leverage for the existing twitter hashtags for building testing and training data where features such as part-of-speech is not suitable for sentiment analysis. It can be evidently seen from the paper [4], that the microblogging features such as hashtags are useful for performing sentiment analysis. The paper is Restricted to lexicon, unigram and bigram features which are the types in N-gram. However microblogging features are well considered for sentiment analysis.

Xujuan Zhou, Xiaohui Tao, Jianming Yong, Zhenyu Yang [5], This IEEE journal paper was published in 2013 projected a novel technique that combines the context-based topic modelling and sentiment analysis which helps in analyzing the opinions of people collected from the microblogging data. The research carried out is on the Australian presidential Election 2010 and is scraped from the Twitter data in the form of tweets. The approach used is the Lexicon based which is doable and is beneficial. The accuracy of the PoS tagging influence overall sentiment. The NLP is being used but can use an advanced version and is limited to a certain NLP technique and emotional analysis. From the results one can observe that modelling reveals topics that are unseen and the tweets are grouped into clusters.

Geetika Gautam, Divakar Yadav, in 2014 IEEE journal paper [7] worked on Sentiment analysis for customer's review classification where machine learning algorithms such as Naïve Bayes, SVM and Maximum entropy. The approach led to the result that Naïve Bayes with unigram gives better accuracy results where the accuracy results for different machine learning algorithms are 83.8% for Maximum entropy, 85.5% for Support Vector Machine (SVM), 88.2% for Naïve Bayes. It was also found that Naïve Bayes works well for large amounts of review data compared to SVM and Maximum Entropy. But the fact that training data set taken is very small and only reviews are considered

André L. F. Alves, Cláudio de S. Baptista, Anderson A. Firmino, Maxwell G. de Oliveira, Anselmo C. de Paiva, wrote a paper [8] in 2014 based on the comparison of SVM and naïve Bayes technique which had been applied on a case study where the Machine learning techniques used were SVM and Naïve Bayes where it was found that both ML approaches are satisfactory in terms of accuracy with SVM accuracy higher than Naïve Bayes for FIFA dataset. It was also evident from the paper [8] that Temporal series can help in the better prediction and a generic solution for all datasets. The results show an accuracy of 72.7% and F-measure of 0.791 for Naive-Bayes sentiment classifier presented and accuracy of 80.0% and F-Measure of 0.873 for SVM sentiments classifier.

Andi Rexha, Mark Kroll, Mauro Dragoni, Roman Kern, in 2016 presented paper [9] in the Semantic sentiment analysis workshop, Greece where Polarity Classification for Target phrases in tweets using the Word2Vec approach was performed the at the have presented automatic prediction of polarity is done in a tweet using Word2Vec model. Word2Vec used for automatically predicting the polarity. SVM and naïve Bayes classification algorithms are used for feature representations. The research findings was that Word2Vec model provided a characteristic spatial representation of words in the training corpus that reflects their relationship to other words. The F1 score measures up to ~90% for the positive class and ~54% for the negative class without using single word polarity information but without exploiting information provided by tweet dependency tresses, classification algorithm performance is reduced. Overall, a good work to predict the target phrase's polarity automatically

III. Proposed System

The system that is proposed and is followed can be seen in the form of a flowchart in Figure 1. Detailed explanation of each step can be seen below:

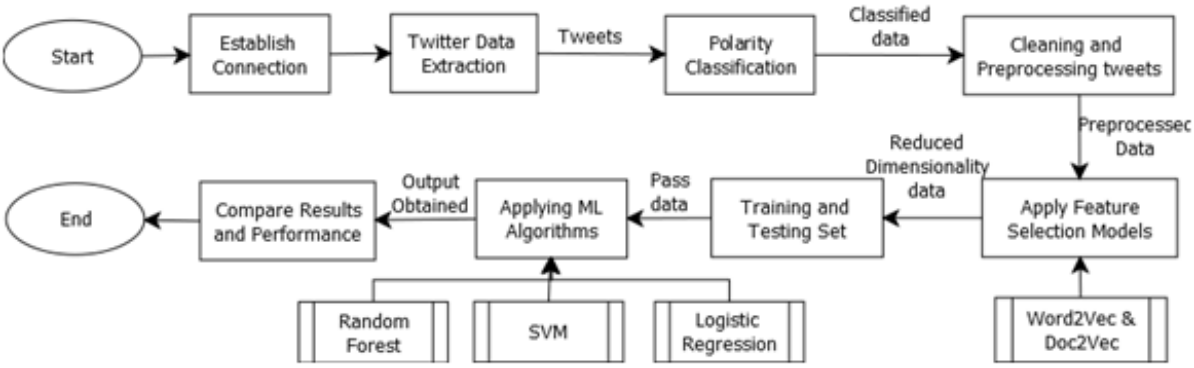

Figure. 1 Flowchart for Proposed System

A. Establishing connection for Tweets Extraction

1. Getting Twitter API keys

Step 1: Login to https://developer.twitter.com/ if already registered else create a new Developer account.

Step 2: After successful login create an Application by filling the form with the required Application details.

Step 3: After successful creation of app open the "Keys and Access Tokens" tab where one can find all the API keys such as Access Token Secret, Consumer Key (API Key), Access Token, Consumer Secret (API Secret)

Step 4: Copy all the API keys required and then use it in the desired programming language for extraction of Tweets from Twitter.

B. Twitter Data Extraction

- Successfully Establishing Connection

After obtaining the API Key, API secret key, Access token and the Access token secret from the developer account, one needs to input these keys in the desired programming language in order to access the Twitter API for data to be extracted.

- Installing Twitter library

By using the required libraries the user can connect to Twitter API and then download the tweets directly from the Twitter through the Twitter API. There are multiple libraries available and supported by most of the programming languages such as R and Python.

- Connecting Programming language with the Twitter API

Initially install the required twitter libraries in order to connect the twitter API with the programming language. After successfully connecting, the tweets can be extracted using a secure connection. For first time extraction one will be directed to Twitter's authorization screen and using PIN one needs to authorize it. Thereby, one can successfully access Twitter API and extract tweets.

- Tweets on #Covid19 can be extracted using the searchTwitter function given the conditions that the extracted tweets are in English and without Retweets.

C. Training data set

Training a data set can be done in several ways where the method here is initially to classify the tweets into the polarity i.e. by performing Polarity Classification where the data set of 200,000 is being split into positive and negative tweets based on the sentiment of each word in the tweet using the sentiment library. The positive indicates a 1 and the negative tweet indicates a 0 which is done by the sentiment library which gives the sentiment values of each of the tweet considering each word and calculating sentiment value and summing up in order to give an accurate sentiment value.

D. Data Preprocessing

Data pre-processing is the process of removing unessential information. It can be achieved in following certain steps as seen in Figure 2 and the steps are explained in detail here:

Step 1: Extracted Tweets are the corpus of data which has been extracted from Twitter using the searchTwitter Function.

Step 2: Check if the Tweets are in English, if Yes proceed to Step 3 else end.

Step 3: Convert all the tweets to lowercase if in case they are in uppercase.

Step 4: Remove the Twitter Handles which is in the form "@[ ]*" so that it can prevent the privacy issues and the GDPR rules are followed.

Step 5: Remove the Punctuations which are comma, full stop, brackets, etc..

Step 6: Remove Numbers and Special characters which are of the form "[0-9]" and "[^a-zA-Z#]" respectively.

Step 7: Remove the Stop words from the tweets and that do not add significant value during analysis. Words such as "the", "an", "a", "his", etc.

Step 8: Normalize the text data which involves reducing the no of unique words in the tweets without losing any relevant or important information.

E. Applying Word Embeddings

- Word2Vec Embedding

Involves a combination of 2 techniques which are Skip-gram model and continuous bag of words (CBOW) where these are shallow neural network techniques basically help in mapping a word to the corpus data which can be a single word or a set of words. The CBOW basically tells the occurrence of a word in the context whereas the Skip-gram model is

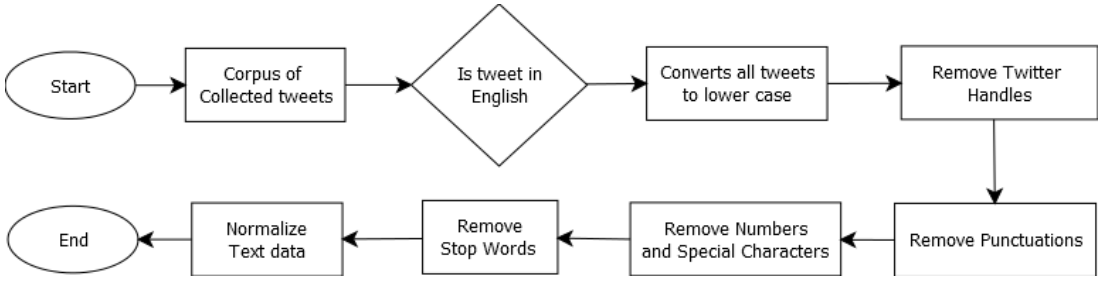

Figure 2 Flowchart for data preprocessing & Cleaning

exactly the reverse of CBOW which tries to predict the context of a given word. The Skip-gram model captures two semantic for a word or words of two vector representations of 'Address' where one can mean to speak to or the other can mean the location. But, the Skip-gram model with the help of negative sub-sampling performs better than the CBOW.

- Doc2Vec Embedding

An extension of the Word2Vec embedding model where the main difference between Word2Vec and Doc2Vec is that Doc2Vec provides a unique additional context for each and every document in the corpus. The other feature is that Doc2Vec provides another feature vector which is the additional content of the whole document. Thereby, the word vectors are trained long with the document vectors.

F. Training and Testing Set

Representations of vectors for all the unique words present in the corpus it is important to training a Word2Vec and a Doc2Vec model on labeled data There is the other choice of using pre-trained vector words rather than training its own model. However, this paper gives an idea of doc2vec and word2vec models being trained instead of the pre trained models as pre-trained vector words size is huge.

The training data from the Word2Vec model shows that the model finds the common or the similar words for a word or a given set of words which indeed acts as a great advantage of this Word2Vec model which helps in identifying similar type of words. One might wonder how the Word2Vec model finds the similar words? The answer to that is that for every different or unique word in the dataset it has a learned vector which uses the cosine similarity in order to identify similar vectors or word.

- Preparing Vectors for Tweets

The Vector representation of the entire tweet dataset is so important that one has to create a way to use the word vector from Word2Vec model since the dataset doesn't just contain words but also a tweet which is a set of words which may have similar set of words or has some meaning. The problem that one needs to create a vector representation can be solved by calculating the mean of all the vector words that is present in the tweet and by assigning the length of the vector to be the same of all the tweets which is considered to be 200. In this same way repeating the process we obtain the vector and also that the training feature vector will also have the Word2Vec feature dataset of 200.

G. Applying Machine Learning Algorithms

- Logistic regression

Logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. It is also a special case of linear regression when the outcome variable is categorical.

- Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm. In this, data items are considered as points in a multi-dimensional space. The classification is performed by constructing the optimal hyper-planes, differentiating the classes well by maximizing the margin between the class's closest points.

- Random Forest

Random Forest is an algorithm for machine learning that perform both regression and classification tasks. It combines a few weak models to form a powerful model. In Random Forest, in contrast to a single tree decision, one grows multiple trees.

IV. ANALYSIS AND RESLT

A. Dataset

The tweets have been extracted for training and testing through twitter developer API using the hashtag Covid19 in English language excluding the retweets. The dataset contains 200,000 tweets and 16 columns which was collected on 26/02/2020 where not much of the outbreak was observed except for china and Italy. The 70% of the total tweets i.e. the dataset which is 1,40,000 is used for training and the rest 30% is used for testing which is about 60,000. The dataset before splitting it to the testing and training dataset, polarity classification is carried out where based on the sentiment value of each tweet, a polarity of positive and negative tweets are assigned where 1 is for positive tweets and 0 is for the negative tweets. From the dataset it has been observed that there are there are 82245 positive tweets and 57755 negative tweets. The testing dataset contains unclassified set of tweets that are used for testing in order to fit into the Machine Learning model.

B. Exploratory Data Analysis

F1 score is used as the evaluation method. Confusion matrix is used to evaluate this model. This matrix contains information about classifications that are predicted and the actual classification as seen in Table 1. It is the weighted average of Precision and Recall.

TABLE1: CONFUSION MATRIX FOR A 2-CLASS CLASSIFIER

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

The important components of F1 score are:

- True Positives (TP) – The positive values when the actual class is yes, and the predicted class is also yes.
- True Negatives (TN) - The negative values when the actual class is no, and the predicted class is also no.
- False Positives (FP) – When actual class is no and predicted class is yes.
- False Negatives (FN) – When actual class is yes but predicted class in no.

$$Precision = TP/TP+FP$$

$$Recall = TP/TP+FN$$

$$F1\ Score = 2(Recall\ Precision) / (Recall + Precision)$$

C. Experimentation Results

1. Construction of Word Cloud

A word cloud represents the importance of a word in the document where the size of the word represents the occurrence or the frequently a particular word as shown in Figure 2 and Figure 3.

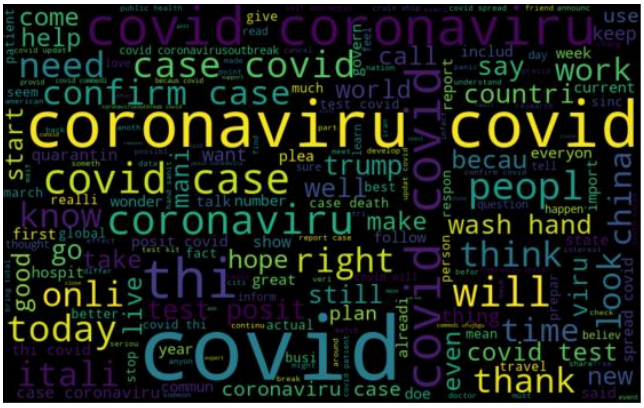

Figure 3 Word Cloud of positive Tweets for #Covid19

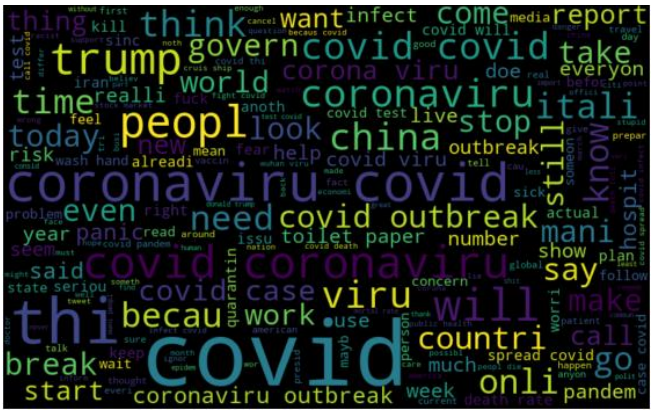

Figure 4 Word Cloud of negative Tweets for #Covid19

The most common tweeted words from the positive and the negative words clouds seen in Figure 3 and Figure 4 are covid, coronavirus, ilali, coronavirusoutbreak, covid case, trump, china, etc. The word cloud from the figures give people's opinions and sentiments which can be related to the ongoing Corona Outbreakthat. These words give us a basic info of what is actually happening around the #covid19 on the social media Twitter. Talking about the observed pattern a on the date the data was scraped, the outbreak was observed most in china and Italy was having its increase which can be observed from the word cloud. This outbreak started in Wuhan, a city in china which until now has taken several lives and is in a state where it's hard to think the stop to this and is yet to officially launch a solution, cure and a vaccine to this scary Virus.

2. Bar Plot Analysis

Analyzing the breakage of categorical data into groups and representing them in the form of graphs is what the bar plot analysis does and can be observed below:

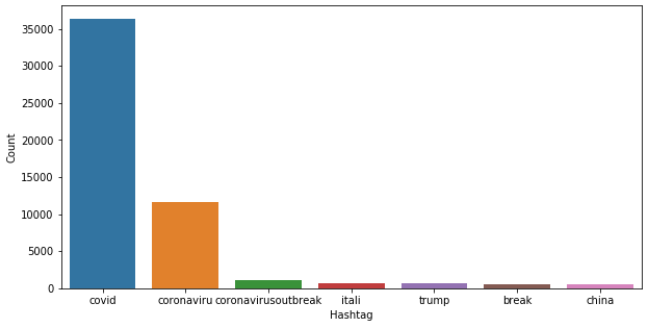

Figure 5 Bar Plot for Positive tweets for #Covid19

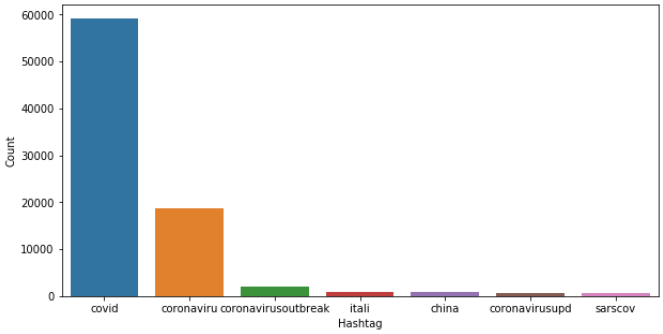

Figure 6 Bar Plot for Negative tweets for #Covid19

The most mentioned hashtag in the 200,000 tweets which were classified as positive and negative tweets can be seen in Figure 5 and Figure 6 respectively. Polarity classification was done in achieving the positive and negative tweets where a sentiment score for each tweet and word was calculated and based on that the positive and the negative most used hashtag are collected and represented in the form of a bar plot. The bar plot is one of the visualization tool that helps in giving a better perspective to a person looking at it. The figure 5 and figure 6 shows the top 7 hashtags of the positive and the negative tweets respectively. The vertical axis represents the total sentiment counts whereas horizontal axis represents seven different Hashtag count where the words are most used.

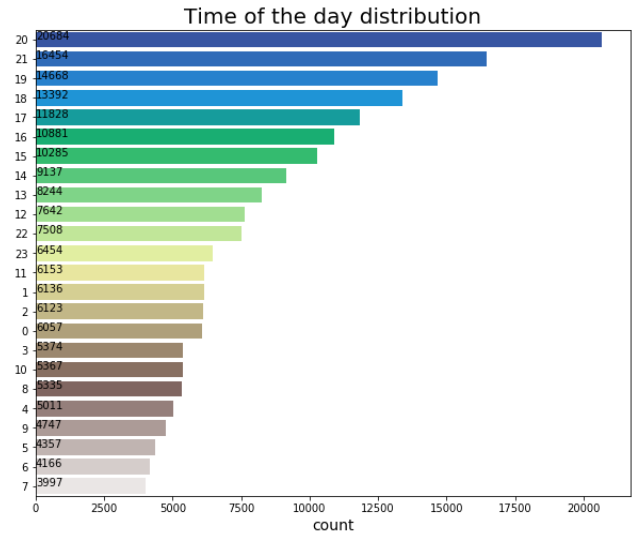

Figure 7 Bar Plot for count at what time most tweets were tweeted

From the Figure 7 one can observe that the at 20th hour of the 24hour day most of the tweets were Tweeted having 20684 tweets and at 7am the least tweets were tweeted having 3997 tweets from the twitter dataset extracted. Data visualization plays important role in summarizing the results in attractive format that is easy to understand and explain the large data set in simple format. In Table 1. Word2vec and Doc2vec feature selection models are compared with two tradition feature selection methods Bag-of-Word and TF-IDF for selecting important feature from tweets and training them to logistic regression, support vector machine and random forest machine learning models for sentiment classification. However, Word2vec gives highest accuracy compared to Doc2vec, BOW and TF-IDF.

TABLE 2: ACCURACY COMPARISION OF DIFFERENT MODELS

	Vector-Space			
Model	**Bag-of-Words**	**TF-IDF**	**Word2Vec**	**Doc2Vec**
Logistic Regression	77.31%	77.24%	77.61%	75.84%
Support Vector Machine	75.58%	76.26%	76.28%	76.18%
Random Forest	75.62%	76.85%	79.29%	75.63%

One can observe from Table 2 that the accuracy of Word2vec embedding model is high when compared with the other models because it uses word embedding which helps in understanding the context of tweets. This paper presents the importance of word embedding by solving Natural language processing (NLP) problems such as sentiment analysis unstructured twitter data.

V. CONCLUSION

In this paper sentiment analysis was performed on #Covid19 Twitter data using a combination of Word Embeddings and Machine Learning model. Initially, twitter developer API account is created and by using the API details and the searchTwitter function 200,000 tweets in English without Retweets on the term #covid19 was collected and classified data into positive and negative tweets using sentiment lexicon. Then data cleaning and data pre-processing is performed on the classified data. Later, Word Embedding feature selection model applied for extracting features from cleaned tweets. Finally, these feature sets are used to build a model for sentiment analysis based on Machine Learning model. The results show that Word2vec with Random Forest for big data improves the accuracy of sentiment analysis by considering contextual semantics of words in the text.

REFERENCES

[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R., 2011. Sentiment analysis of twitter data. In Proceedings of the Workshop on Language in Social Media (LSM 2011) (pp. 30-38).

[2] Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S., 2012, July. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the ACL 2012 System Demonstrations (pp. 115-120). Association for Computational Linguistics.

[3] Liu, B. (2015) "Preface," in Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge: Cambridge University Press, pp. xi-xiv. doi: 10.1017/CBO9781139084789.001.

[4] Kouloumpis, E., Wilson, T. and Moore, J., 2011, July. Twitter sentiment analysis: The good the bad and the omg!. In Fifth International AAAI conference on weblogs and social media.

[5] Zhou, X., Tao, X., Rahman, M. and Zhang, J. (2017). Coupling topic modelling in opinion mining for social media analysis. Proceedings of the International Conference on Web Intelligence - WI '17.

[6] Neethu, M.S. and Rajasree, R., 2013, July. Sentiment analysis in twitter using machine learning techniques. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[7] Gautam, G. and Yadav, D., 2014, August. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In 2014 Seventh International Conference on Contemporary Computing (IC3) (pp. 437-442). IEEE.

[8] Firmino Alves, A., Baptista, C., Firmino, A., Oliveira, M. and Paiva, A. (2014). A Comparison of SVM Versus Naive-Bayes Techniques for Sentiment Analysis in Tweets.

[9] Rexha, A., Kröll, M., Dragoni, M. and Kern, R. (2016). Polarity Classification for Target Phrases in Tweets: A Word2Vec Approach. The Semantic Web, pp.217-223.

[10] Noferesti, S. and Shamsfard, M., 2015. Using Linked Data for polarity classification of patients' experiences. Journal of biomedical informatics, 57, pp.6-19.

[11] Kuamri, S. and Babu, C.N., 2017, July. Real time analysis of social media data to understand people emotions towards national parties. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[12] Chen, Q. and Sokolova, M., 2018. Word2Vec and Doc2Vec in Unsupervised Sentiment Analysis of Clinical Discharge Summaries. arXiv preprint arXiv:1805.00352.

[13] Haripriya, A. and Kumari, S., 2017, July. Real time analysis of top trending event on Twitter: Lexicon based approach. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-4). IEEE.

[14] Hughes, M., Li, I., Kotoulas, S. and Suzumura, T., 2017. Medical text classification using convolutional neural networks. Stud Health Technol Inform, 235, pp.246-250.

[15] Lau, J.H. and Baldwin, T., 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. arXiv preprint arXiv:1607.05368

[16] M. Hitesh, V. Vaibhav, Y. J. A. Kalki, S. H. Kamtam and S. Kumari, "Real-Time Sentiment Analysis of 2019 Election Tweets using Word2vec and Random Forest Model," 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, India, 2019, pp. 146-151.