How to Cite:
Padmavathi, J., & Raja, V. (2022). A study on the impact of Corona Virus and it's mutants in India using machine learning algorithms. *International Journal of Health Sciences*, *6*(S3), 4134–4145. https://doi.org/10.53730/ijhs.v6nS3.6748

A study on the impact of Corona Virus and it's mutants in India using machine learning algorithms

Dr. J. Padmavathi
Associate Professor, Department of Computer Science & Application, SRMIST, Chennai, India

Raja V.
Assistant Professor, Department of Computer Science & Application, SRMIST, Chennai, India

***Abstract*---**With the advent growth of technology, large volumes of data are available in the internet for researchers to explore. Social Media is one such platform that helps researchers in analyzing various data for various reasons such as improved Customer Service, development of quality Products. , discovering New Marketing Strategies, improve Media Perceptions and much more. The Social networking sites such as Facebook, Twitter, Instagram are being used by people on a large scale to share and express their views in the form of text, emoji and post. This data is used for sentiment analysis. This paper aims at analyzing the human sentiments and emotions in the second wave of Corona virus in India and about the awareness of vaccination using twitter data. Machine learning algorithms such as Naïve Bayes, Logistic Regression and Support Vector Machine were implemented on the twitter data set. As the pandemic situation has created an alarming situation with new symptoms, the disease has affected many in India. This study will assist the government agencies and health care volunteers to better assess the mental state of public and to bring more awareness on safe and secured living by adapting precautionary measures.

***Keywords*---**corona virus, logistic regression, machine learning algorithms, naïve bayes, sentiment analysis , social media, support vector machine, twitter.

Introduction

The tremendous use of social media has contributed enormous data for research. It is observed that Indians on average spend more than 2 hours on social media daily. Due to the availability of internet connectivity the number of social media users in India is about 448 million. This is roughly about 45% of the total population of India. 50.60% of people use twitter now both in Mobile and in Computers. During the second wave of COVID-19, India faced severe consequences in the form of multiple complications in treating Covid patients, insufficient supplies of essentials like oxygen cylinders, drugs for the treatment and increased death rate; particularly in the young population. It is very challenging for the researchers and the doctors to identify the control measures as the double-mutant and triple mutant of the virus had very quick transmissibility ending up in different symptoms. It was observed that in the first wave, senior citizens were badly affected and in the second wave young people aged between 25 to 45 years were most affected. The death rate was very high creating a panic situation among the people. Further people who were treated for Coronavirus had black fungus attack. As of June 7, 2021, the Indian Ministry of Health had recorded 28252 cases of black fungus. The extent use of twitter paved way for performing Sentiment analysis in recent days to predict and/or monitor business, Political situations and health related issues. As the tweets are of short lengths, prejudiced or irrelevant messages or polluted content, which creates a negative effect in perception analysis. The research makes use of spatio-temporal reasoning of twitter data by searching tweets by location and with specific keywords in the pandemic situation.

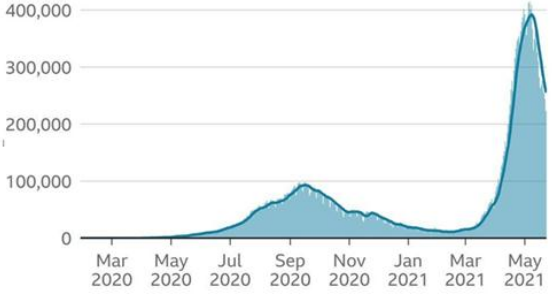

Fig 1. Growth of New Cases in India

Literature review

Er. Puneet et al, have vividly summarized the usage of twitter data for sentiment analysis and the various tools required to perform the different steps. KavyaSuppala and Narasingarao, have showed that that customer opinion and perceptions can be enhanced to desired level by using Naïve Bayes algorithm. The study was to perform analysis on tweets having sentiment which help in business intelligence on predicting the future. Rakshitha C L and Gowrishankar S, have analysed the mental health and wellbeing of a person using machine learning algorithms. The emotional keywords used by the individuals are used to identify their stable mind condition and Naïve Bayes classifier is used for classification.

Zulfadzli Drus, Haliyana Khalid, have reviewed on articles published from 2014- to 2019 on the application of Sentiment analysis in Social media. They have recorded that the utmost application of opinion-lexicon method is widely used to analyze text sentiments on Twitter data on world events, healthcare, politics and business. Agarwal et al, have applied ConceptNet to construct a domain specific ontology for product reviews and, WordNet is used to expand the ontology for better coverage of the product features. The effects of domain specific ontology, importance of the features, and contextual information are included in determining the overall sentiment of text. The authors have concluded that, along with ConceptNet, other ontologies can help also to enrich the concept mining process at a great level in the near future.

Shyed , et al, explained that , Sentiment analysis is an analysis of subjective word in the text which provides user's opinion towards entities on social media websites. In this paper, various techniques for sentiment classification are discussed with their pros and cons. Ra ahuja, have implemented 6 different classification algorithms on the SS-Tweet dataset with two features (TF-IDF and N-Grams). It was shown that, TF-IDF features are giving better results (3-4%) when compared to N-Gram features. Prima Widyaningrum et al, described the twitter sentiment analysis for "trust", and "anticipation" of public sentiment of a company using Chatbot technology using R- Studios. Razia Sulthana et al**,** have applied data analytic approach for predictive modeling on Hilary Trump data set and found that the method obtained high accuracy than support vector machine and Naïve Bayes approach.

Ajay Bandi and Aziz Fellah, have proved that the precision values of Socio-Analyzer and TextBlob are 70.74% and 72.92%, respectively, when considered neutral tweets as positive. Bryan Pratama et al, stated that the qualitative testing process, classifies text with high accuracy, while Part-of-Speech (POS) is used to assign a tag known as word class grammatically to each word in a text sentence using Rapidminer. Alaa Khudhair Abbas et al, have implemented ensemble majority Vote classifier to categorize each tweet sentence class. The base classifiers NB, LR, MLP and DT have the same weight in the process. The classification training is implemented using Weka and LightSIDE for feature vector extraction and data cleansing. The authors have shown clear cut evidence on the high classification rate (of 82.6%) by Ensembled Majority vote unigram features followed by the NB base classifier with a classification rate of 80.05%.

Sentiment analysis and machine learning

Sentiment analysis is approach that uses Natural Language Processing (NLP) to extract, convert and interpret opinion from a text and classify them into positive, negative or natural sentiment [4]. The basic idea of sentiment analysis is to detect the polarity of text documents or short sentences and classify them correctly. The most popular methods used in sentiment analysis are:

- Lexicon based Approach
- Machine Learning Approach

Lexcion based approach: Lexicon-based approach works by counting the positive and negative words that related to the data. Under this approach lexicons are grouped to form a collection called dictionary. In Corpus based approach, the occurrences of the term with other positive or negative seed words in the corpus is used to compute the polarity value of a term. The dictionary based approach makes use of the pre-developed polarity lexicons like SentiWordNet, WordNet, SenticNet and so forth. Researchers have experimented with many methods but still, the most common method used for Lexicon based approach includes SentiWordnet [5] and TF-IDF. On the other hand Machine learning approach uses Naïve Bayes, Logistic Regression and SVM. The choice of appropriate method for sentiment analysis is data dependent.

Model construction

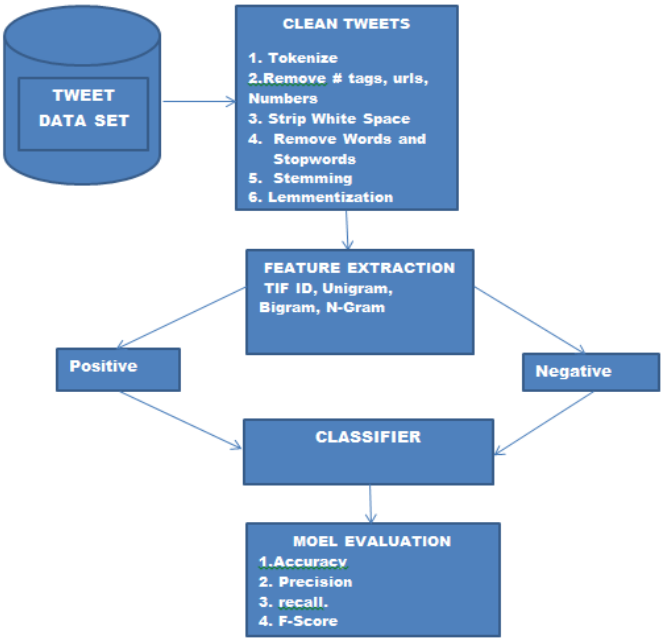

Fig 2. Classifier Model For Tweet Analysis

Algorithm

- Create a new twitter account and log in with the username and password.
- Create a new app and fill in the credentials agreeing to the terms and create the twitter application.
- Extract the "API keys" and "API secret" information such as consumer key, consumer secret, access key, access secret.
- Select and copy the "Access token" and "Access token secret".
- Eliminate all URLs (e.g. www.xyz.com), hash tags (e.g. #topic, targets (@username)
  - Make Corrections in the spellings; sequence of repeated characters is to be handled
  - Find all the emoticons and replace with their respective sentiments.

- Eliminate all punctuations ,symbols, numbers
- Eliminate Stop Words
- Expand Acronyms (we can use an acronym dictionary)
- Eliminate Non-English Tweets

- Build the tweet Dictionary and compute polarity
- Perform feature selection based on polarity. Use Unigram, bigram, N-gram or Term Frequency-Inverse Document Frequency (TF-IDF) method.
- Build Classifier Model 'M'.
- Apply Training data set to label as Positive, Negative or Neutral
- Evaluate 'M' with test data set.

Naïve bayes classification algorithm

Naïve Bayes is also a classification algorithm that is based on the principle of Bayes Theorem. It follows Bayes theorem. It is a probabilistic model and makes use of conditional probability for prediction. The dataset is subdivided into feature set and response vector. The feature set has dependent features and the response vector has predictor or the class label. The assumption is that each feature in the dataset is independent of each other and renders equal contribution to the result. Mathematically, Conditional probability of A given B can be computed as:

$$P\left(C_k/_x\right) = \frac{P(C_k)P(x/_{C_k})}{P(x)}$$

Where:

$P(C_k|x)$: Probability (conditional probability) of occurrence of event C_k given the event 'x' is true. $P(C_k)$ and P(x): Probabilities of the occurrence of event C_k and x respectively.
$P(x|C_k)$: Probability of the occurrence of event 'x' given the event C_k is true.

Logistic regression algorithm

Logistic regression is a linear algorithm that takes binary input set and performs non-linear transformation on output. The basic assumption is, there exists a linear relationship between the input variables and the output. As a result of output data transformation the model produces more accuracy. It is much preferred in situations in which we would want to predict the presence or absence of a characteristic based on values of a set of predictor variables. Logistic regression is used to analyze relationships and metrics between the binary dependent or binary independent variables. Logistic regression combines the independent variables to estimate the probability that a particular event will occur. In other words, it ascertains that a subject will be a member of one of the groups defined by the binary dependent variable. The output value produce by LR is a probability within the range 0.0 or 1.0. A cutoff value; say 0.5, is fixed to classify if a data point is a member of the group or not. Logistic regression makes use of the sigmoid function which produces a probability between 0 and 1.

$logit(p) = \log(\frac{p}{1-p})$, where 'p' is the probability and '1-p' is the corresponding odds.

Support vector machine

The Support Vector Machine acts as a binary classifier. SVM calculates the hyperplane with largest margin and uses it to separate dataset into two different classes. The two types of separations permissible in SVM are linear and non-linear. In linear SVM, the classifier separates the dataset using a straight line. The non-linear SVM, the classifier works with kernel functions to separate data set in two- dimensional space and further lifts it to higher dimensions such as three-dimensional space thereby creating hyperplanes. The classifier separates the data points with maximum margins. The kernel function used in this study is Gaussian Kernel Function as it needs no prior knowledge about the data.

Proposed work

The proposed work was to analyze the performance of machine learning algorithms, namely the Naïve Bayes Classifier and Support Vector Machine on Twitter data. The twitter data was collected using Python API from 1st June to 3rd June 2021 with geological restrictions. This data was cleaned and classifier was built using MonkeyLearn. MonkeyLearn is a machine learning platform that makes it easy for the user to build and implement sentiment analysis. It permits to build and train the model using our own Twitter data. The performance of the model can be evaluated using the test data with confidence measure as performance parameter. The Naïve Bayes classifier was designed using MonkeyLearn API and the dataset was fed into the classifier as CSV file or as Excel. Duplicate values are removed and the dataset is imported as shown in Figure 3. In the training phase the data is tagged as positive, negative or neutral for fewer examples or for the entire set. After building the Model, the classifier is tested with batch file and the confidence values for the test data can be downloaded as shown in figure 3 and figure4. We can connect the models with our Apps like Rapidminner, Zapier, Googlesheets, etc,. In our research, MonkeyLaern API using Python is used to build the Navie Bayes classifier. The classifier partitions the data into train set and test. The accuracy of the model was nearly 78%.

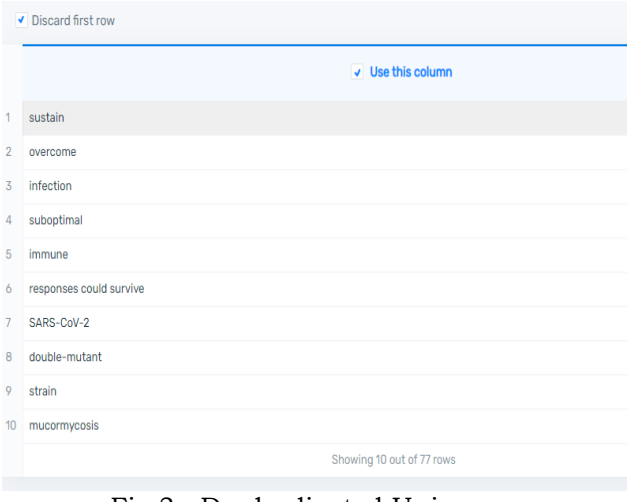

Fig 3. De-duplicated Unigrams

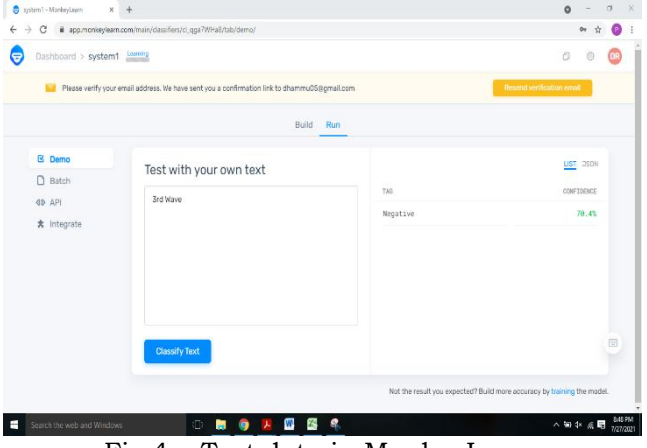

Fig 4a. Test data in MonkeyLearn

Tweets	Tags	Classification	Confidence
3rd wave	negative	Negative	0.704
children	neutral	Negative	0.671
Death	negative	Negative	0.406
fungal infections	negative	Negative	0.7
Transmissible	negative	Positive	0.374
virus variant	negative	Negative	0.466
Mutating virus	negative	Negative	0.453
Undercounting	negative	Positive	0.372
deaths	negative	Negative	0.415
Statesï¿½	neutral	Positive	0.389
interventionspay	positive	Positive	0.373
Policy flip-flops	negative	Negative	0.458
vaccination	positive	Negative	0.456
think	positive	Negative	0.384

Fig 4b. Processed Batch file with Confidence values

sars-cov2	16
black fungus	18
white fungus	3
vaccinate	20
think	10
lockdown	13
virus	12

Fig. 5. Polarity of Unigrams and Bigrams

Fig 6. Word Cloud

Experiment results

Twitter data set was collected using Twitter API from 1st June to 3rd June. The data was fed into Python and preprocessing and cleaning operations were performed. This processed dataset was analyzed using the popular Machine learning algorithms namely Naïve Bayes, Logistic Regression and Support vector machine. Unigram Model. It is proved by many researchers that machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) has yielded appreciable amount of success in sentiment analysis domain [12]. The unigram language model makes the following assumptions:

- The probability of each word is independent of its preceding words appearing in the dataset.
- The fractions for all unigrams generated using the training text dataset is used in training the model
- i.e. $$P_{train}(word_1) = \frac{n_{train}(word_1)}{N_{train}}$$

And N_{train} is the total number of words in training text.

- Evaluating the model is done using the Average log likelihood of the test data or the evaluation text.

$$P_{eval}(text) = \prod_{word} P_{train}(word)$$

$$\log(P_{eval}(text)) = \sum_{word} \log(P_{train}(word))$$

$$\text{Average log likelihood}_{eval} = \frac{\sum_{word} \log(P_{train}(word))}{N_{eval}}$$

Where N_{eval} refers tothe total number of words in evaluation text.

Bigram model

In Bigram model the probability of a word is calculated by using the conditional probability of the one previous word. i.e.

$$P\ (Word_n | Word_{n-1})$$

Maximum Likelihood Estimation (MLE) is used estimate the evaluation text. For example, to compute a particular bigram, probability of a word *'y' along with the occurrence of its* previous word *'x'*, is determined by using the count function bigram C(x, y) and this computed value is normalized by summing all the bigrams that share the same common first-word *'x'*.

TIF-IDF method

The term frequency is a weight value assigned to the words based on their frequency of occurrence at any given time. The Term Frequency Inverse Document Frequency (TF-IDF) is executed to observe the weight values of a term against a positive or negative sentiment registered by the user. The formula is

$$tfid(word) = count_{t(word)} x \log(N\ x\ count_{d(word)})$$

Where:

- $count_t$ = the number of occurrences of term word in the document
- $count_d$ = the number of occurrence of documents containing the term word
- N= Total number of documents

Evaluation of results

The twitter data set was downloaded and the analysis feature selection was done using Unigram, Bigram method and TF-IDF method was used.
The performance of sentiment classification can be evaluated by using four indexes calculated as the following equations: Accuracy = (TP+TN) / (TP+TN+FP+FN)
Where:

- TP -number of true positive instances
- FN - number of false negative instances
- FP - number of false positive instances
- TN - number of true negative instances

The obtained results are tabulated as shown in Table1. It was observed that Navie Bayes Classifier performed better than Logistic regression and SVM classified the tweets more accurately.

Table 1
Performance of the Classifiers using various Models

CLASSIFIER MODEL	ACCURACY (%) UNIGRAM	ACCURACY (%) BIGRAM	ACCURACY (%) TIF_ID
Naïve Bayes (NB)	77.551	73.23	76
Logistic Regression (LR)	73.4	75.6	75.8
Support Vector Machine (SVM)	79.5	80.5	82.3

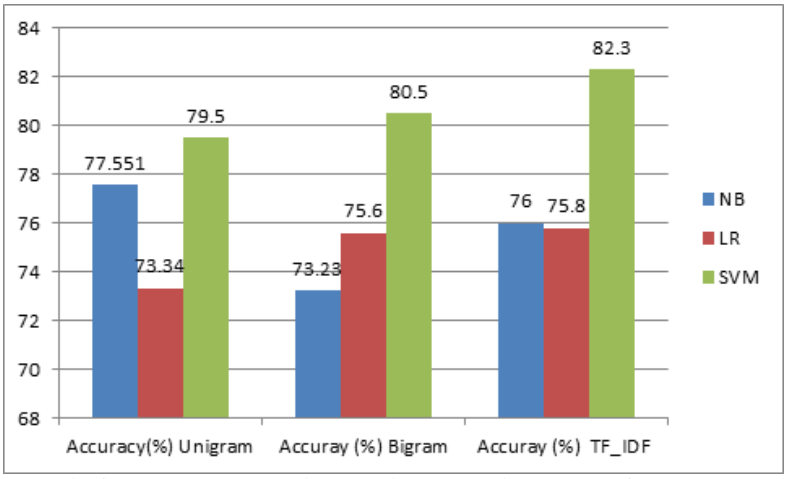

Fig 7. Accuracy using Unigram, Bigram and TF-IDF

Discussions

The findings of some of the authors in Machine learning algorithms with Twitter dataset is compared with our working. Rakshitha C L and Gowrishankar S have applied Naïve Bayes algorithm and SVM for text classification in Mental illness and have observed that in weekend the negative tweets were more than the positive tweets. The author has not recorded the metrics. KB Priya Iyer & Sakthi Kumaresh have shown 70% accuracy in classifying tweets under Naïve Bayes , where as our classification algorithm has produced 77.5% accuracy using Naïve Bayes Classifier. Kavya Suppala, Narasinga Rao in Sentiment Analysis Using Naïve Bayes Classifier has used Bag of words concept that contains both positive and negative tweets and accuracy of the model is not recorded. Abdullah

Alsaeedi1, Mohammad Zubair Khan have recorded that the Naive Bayes, Maximum Entropy, and SVM, achieved an accuracy of approximately 80% when n-gram and bigram model were utilized. Ensemble and hybrid-based Twitter sentiment analysis algorithms tended to perform better than supervised machine learning techniques, as they were able to achieve a classification accuracy of approximately 85%.

- Alaa Khudhair Abbas, Ali Khalil Salih, Harith A. Hussein , Qasim Mohammed Hussein, Saba
- Alaa Abdulwahhab, Twitter Sentiment Analysis Using An Ensemble Majority Vote Classifier, NB using Unigram was 80%, Bigram was 80.21% and tri-gram was 63.85%.

Conclusion

The aim was to analyze the efficiency of some of the machine learning algorithms in classifying the tweets. As there is enormous contributions made by the researchers in sentiment analysis it was further decided to examine the contribution of the various feature extraction methods in the classification process. It was observed that Naïve Bayes classifier performed better with TF-IDF method when compared to Unigram and Bigram methods for feature extraction. Logistic regression was applied to text classification and the performance was appreciable. Among the three machine learning algorithms, SVM produced the highest accuracy of 82.3% with TF-IDF method in classifying Twitter data. The study analyzed the mindset of the public during pandemic situation. It confirmed that people from all parts of the country have understood the need for vaccination and have inculcated the sense of responsibility in fighting against novel corona virus and its mutants.

References

1. Er. Puneet, Dr. Vinay Goyal, Er. Rajdeep Kaur**,** Survey: Sentiment Analysis on Twitter Data using machine learning classification techniques, International Journal of Information and Computing Science, Volume 5, Issue 6, June 2018.
2. KavyaSuppala and Narasingarao , “Sentiment analysis using Naïve Bayes classifiers”, International Journal of Innovative Technology and Exploring Engineering, June 2019
3. Rakshitha C L and Gowrishankar S, “Machine Learning based Analysis of Twitter Data to Determine a Person's Mental Health Intuitive Wellbeing”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 21 (2018)
4. Zulfadzli Drus, Haliyana Khalid, “ Sentiment Analysis in Social Media and Its Application: Systematic Literature Review”, www.sciencedirect.com Procedia Computer Science 161 (2019) 707– 714.
5. Agarwal, Basant, Namita Mittal, Pooja Bansal, and Sonal Garg. (2015) “Sentiment Analysis Using Common- Sense and Context Information.” Journal of Computational Intelligence and Neuroscience 9 (2015).
6. Ebrahimi, M.m Yazdavar, A., and A. Sheth. (2017) “On the Challenges of Sentiment Analysis for Dynamic Events.” Intelligent Systems, IEEE 32 (5).

7. Ravinder Ahuja, Aakrasha Chug, Shruti Kohli, Shaurya Gupta, Pratyush Ahuja," The Impact of Features Extraction on the Sentiment Analysis" Procedia Computer Science. 152. 341-348.10.1016/j.procs.2019.05.008. , (2019).
8. Syed Saood Zia, Sana Fatima, IdrisMala, M. Sadiq Ali, Khan, M. Naseem, Bhagwan Das, "A Survey on Sentiment Analysis, Classification and Applications", International Journal of Pure and Applied Mathematics,Volume 119 No. 10 2018, 1203-1211.
9. Prima Widyaningrum, Yova Ruldeviyani, Ramanti Dharayani , "Sentiment Analysis to Assess the Community's Enthusiasm Towards the Development Chatbot Using an Appraisal Theory", The Fifth Information Systems International Conference , Procedia Computer Science. (2019).
10. A Razia Sulthana, A K Jaithunbi, L Sai Ramesh, "Sentiment analysis in twitter data using data analytic techniques for predictive modeling", National Conference on Mathematical Techniques and its Applications (NCMTA 18) , IOP Publishing, IOP Conf. Series: Journal of Physics: Conf. Series (2018).
11. Vishal A. Kharde, S.S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications (0975 – 8887),Volume 139 – No.11, April 2016.
12. Ajay Bandi and Aziz Fellah, "Socio-Analyzer: A Sentiment Analysis Using Social Media Data", EPiC Series in Computing, Volume 64, 2019, Pages 61-67.
13. Bryan Pratama, Dedi Dwi Saputra, Deny Novianti, Endah Putri Purnamasari, Antonius Yadi Kuntoro, Windu
14. Gata, Nia K Wardhani, Sfenrianto Sfenrianto, Sularso Budilaksono, "Sentiment Analysis of the Indonesian Police Mobile Brigade Corps Based on Twitter Posts Using the SVM and B Methods", Journal of Physics: Conference Series, volume 1201, IOP Publishing, (2019).
15. Alaa Khudhair Abbas , Ali Khalil Salih , Harith A. Hussein , Qasim Mohammed Hussein , Saba Alaa
16. Abdulwahhab, "Twitter Sentiment Analysis Using An Ensemble Majority Vote Classifier", Journal of Southwest Jiaotong University, Vol. 55 No. 1, Feb. 2020.
17. Abdullah Alsaeedi1, Mohammad Zubair Khan, "A Study on Sentiment Analysis Techniques of Twitter Data", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 10, No. 2, 2019.
18. K B Priya Iyer, Sakthi Kumaresh, "Twitter Sentiment Analysis On Coronavirus Outbreak Using Machine Learning Algorithms", European Journal of Molecular & Clinical Medicine , volume 07, Issue 03, 2020.