International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

Naïve Bayes Classification of Sentiments on Subset using Tweets-during Covid-19

V.Geetha[1,2*], N.Sujatha[3], Latha Narayanan Valli[4]

Submitted: 05/02/2024 **Revised**: 13/03/2024 **Accepted**: 19/03/2024

Abstract: Now a day, Social Media create a platform for almost all people for sharing and communicating with one another. Most of the business people and the organizations avail the social media conversation for their product promotion or predicting people behavior. The popular Social Media Networks are Facebook, Twitter, LinkedIn of Social Networks, Instagram, YouTube of Media Sharing Networks, Whatsapp, Pinterest and tripAdvisor of Consumer Review Networks. A Text Mining tool, Sentiment Analysis can help us to predict and classify the susceptible text used in the social media conversation. Even though having lots of advantages, unfortunately we have many risks in the usage of the social media content. Any individual must follow the rules and regulations for accessing the content in the social media networks. The objective of this research paper is to understand the various techniques involved in Sentiment Analysis process and choose to apply naïve bayes machine learning model in the subset level using twitter data to classify the sentiments of people in a best way.

Key Words: *Social Media Data, Twitter, Sentiment Analysis, Machine Learning Algorithms, naïve Bayes Algorithm, Text Processing*

1. Introduction

The social media chat can be a database for content analysis. We can apply machine learning algorithms to classify the text and predict the behavior of an individual based on the constraint set. Sentiment analysis is a Text Mining tool for understanding emotions and behaviors of an individual.

The supervised and unsupervised machine or deep learning algorithms can support different professionals to understand the attitude to take decisions to support any individual in a particular situation. But privacy is a major concern when we are using social media content [1].

1.1 Social Media Data

Social Media is well-liked to convey any type of information very quickly at any time. The huge data is formed in social media at every second. The quality and effectiveness are so important in any communication. We can perform text classification based on user's reviews and opinions. The content can be motivating, discouraging, relevant or irrelevant to the particular group [2].

1.2 Sentiment Analysis

The sentiment analysis in text mining involves discerning subjective (as opposed to factual) material and also extracting attitudinal information like sentiment, opinion, mood, and emotion. By machine learning algorithms, we can classify the words such as fear, anxiety, happy and so on. Text analytics techniques are helpful in analyzing sentiment at the entity, concept, or topic level [3].

1.3 Risks using Social Media Data

Social media generates both qualitative as well as quantitative data in every minute. People were interacting with each other either knowingly or unknowingly about the practices and guidelines of social media. Any researcher, who had primary or secondary social media data, should consider the type of data, the techniques used for data collection and the methodologies to adopt throughout the research process in order to maintain the research quality [4].

Many people involved in tweet by using the popular social media, Twitter. Now a day, researchers from various domains were using tweets for analysis. The twitter user may or may not know about the researcher and the type of research undergone by them. So the researchers should think in user point of view while using data by following the guidelines which were provided by the association like Association of Internet Researches [5].

2. Related Work

The structure, length and language of the text used in various social media conversation were different. The method for the feature selection and supervised machine learning algorithm can be used for detection and classification of interesting pattern, since the communication is not same in printing or non-printing media [6]. The feature selection plays an

[1]*Research Scholar, P.G & Research Department of Computer Science, Sri Meenakshi Govt. Arts College for Women (Autonomous), Madurai Kamaraj University, Madurai, Tamil Nadu , India*
[2]*Assistant Professor,*
Department of Computer Science,
Lady Doak College,
Madurai Kamaraj University, Madurai
Tamilnadu, India
Email: geetha@ldc.edu.in
[3]*Associate Professor,*
P.G & Research Department of Computer Science,
Sri Meenakshi Govt. Arts College for Women (Autonomous),Madurai Kamaraj University, Madurai
Tamilnadu , India
Email:sujamurugan@gmail.com
[4]*Vice President,*
Standard Chartered Global Business Services Sdn Bhd,
Kuala Lumpur, Malaysia
Email:latha.nv@gmail.com

important role in achieving a greater accuracy of the result, if there were number of features in the data set [7].

The predictions of people's sentiment in different domains were very useful to improve or increase the market levels for business organizations. For example, the sentiment analysis can be applied to predict stock level, financial market, election forecasting, movie reviews, web care, and online reputation monitoring [8].

The framework of research ethics must guarantee for the rights and welfare of both people and data. Whatever methodology was used by the researcher, the ethics should covers the privacy of an individual, maintain the confidentiality of information, the authorization of data and minimize the harm for people as well as data. There were number of global bodies exist for providing guidelines for follow ethics [9].

3. Machine Learning Algorithms

Using machine learning algorithms, the data can be transformed into intelligent action. The algorithms accept the data and identify the interesting patterns and change that into actions. The machine learning process has three components such as Data input, Abstraction and Generalization. The various machine learning models and few of their applications are given below [10]:

- Nearest Neighbors: Recognition of character and face in images and video files, Predicting movie reviews, Pattern identification in sequences like proteins
- Naïve Bayes: Classification of text, Detection of intrusion in computer networks, Diagnose medical conditions
- Decision Trees, Classification Rule Learner: Useful in credit scoring systems, Finding customer satisfaction in advertisement field, Diagnose medical conditions
- Linear Regression, Regression Trees, Model Trees: Finding characteristics of an individual or group, predicting causal relationship between variable, Recognize patterns for future prediction
- Neural Networks, Support Vector Machines: Recognize speech and handwriting, Automating smart devices, Build sophisticated models for weather and climate patterns
- Association Rules: Finding interesting patterns in DNA and protein sequences, Fraud detection in purchase or medical claims, Searching human behaviors
- Clustering with k-means: Predicting purchasing behavior of customers, finding unauthorized users in computer networks, Detecting small number of homogeneous group from larger data sets

4. Different Social Media Data And Its Applications

The observed resultant from some of the social media data, algorithm and the corresponding purpose are shown in the following table:

Table 1. Data Set, Algorithm Used

Author Name &Year	Data Set Used	Purpose	Algorithm Applied	Observations
[Widiastuti N.I., 2018] [11]	Twitter-Word and additional information considered	Identification of network performance	Utilizing hidden layer for feature selection	Recognize the context of sentences in documents
[Sahar Sohangir et al. 2018] [12]	Stock Twits	improve the accuracy of sentiment analysis of Stock Twits messages	RNN, CNN	Doc2vec, LSTM,CNN were applied to find the accuracy in those who read messages in Stock Twits
[Ha Sung Hwang 2019] [13]	Instagram-focused on photos	Upward, Downward prediction of social comparison	Survey Method, SPSS	Comparison of depressive mood of people
[Concetta Papapicco 2019] [14]	Whatsapp	Organizational well-being, support to others	qualitative-quantitative methods	Speed up decision making and increase positivity in working groups
[Chinthapanti Bharath Sai Reddy et al. 2020] [15]	Whatsapp	Predicting emotions from the text	long short term memory, bag of words	Maintaining the position with other person in the chat by using proper words
[Abubakar Ahmad et al. 2021] [16]	Whatsapp	Investigation of people's involvement in the group	Python libraries such as Numpy, Pandas, Matplotlib,	Find people participation in the group

5. Methodology And Application Using R Tool

Based on feature values, an observed probability of each class is computed using the training data. The application of prior knowledge to the probability of an outcome, thereby met the higher probability for the actual outcome. The Bayes' probability rule is given below:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)} \quad (1)$$

- Where A and B are conditionally related events and P(A|B) denotes the probability of event A occurring when B has already occurred

One of the most important sources of information is textual data. It is inherently unstructured. Identifying human feelings from social media networks of data is one application of text analytics. The most crucial step in creating a model is preprocessing because the text is in an unstructured format [17]. The following are the steps in naïve Bayes text classification:

5.1 Loading the Dataset

The dataset used to perform text categorization was obtained from Kaggle Open Source in 2020 [18]. After the tweets were taken down from Twitter, manual tagging was completed. To allay any worries about privacy, distinct codes have been assigned to the identities and names. Location, tweet at, original tweet, and sentiment are the remaining features. The model's dataset size, as indicated below:

Table 2. Twitter Dataset

No of Rows	No of Columns
41157	6

The following table displays the first three records from the dataset:

Table 3. First Three Records of loaded data

A data.frame: 3 × 6						
	UserName	**ScreenName**	**Location**	**TweetAt**	**OriginalTweet**	**Sentiment**
	<int>	**<int>**	**<chr>**	**<chr>**	**<chr>**	**<chr>**
1	3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/iFz9FAn2Pa and https://t.co/xX6ghGFzCC and https://t.co/I2NlzdxNo8	Neutral
2	3800	48752	UK	16-03-2020	advice Talk to your neighbours family to exchange phone numbers create contact list with phone numbers of neighbours schools employer chemist GP set up online shopping accounts if poss adequate supplies of regular meds but not over order	Positive
3	3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elderly, disabled dedicated shopping hours amid COVID-19 outbreak https://t.co/bInCA9Vp8P	Positive

5.2 Exploring the Dataset

There are 41157 records in the dataset, and each record has five manually categorized sentiment levels. This study work's goal was to identify and select a subset, namely a specific day, for classification. The following tables provide the data, five levels, and sentiment count:

Table 4. Information of the dataset

```
'data.frame': 1428 obs. of  6 variables:
 $ UserName     : int  43244 43245 43246 43247 43248 43249 43250 43251 43252 43253 ...
 $ ScreenName   : int  88196 88197 88198 88199 88200 88201 88202 88203 88204 88205 ...
 $ Location     : chr  "Kolkata/Delhi" "Kenya " "Hou" "Copperbelt, Zambia" ...
 $ TweetAt      : chr  "13-04-2020" "13-04-2020" "13-04-2020" "13-04-2020" ...
 $ OriginalTweet: chr  "It may take 9 months for the economy to recover from the 19 shock says Bank Chairman Deepak Parekh He warns pri"| __truncated__ "So Brookside buys 1litre of milk from farmers at about 35 removes fat to make ghee butter and gives you back wa"| __truncated__ "To health workers To security officers To artists To teachers To scientists To grocery store workers To food pr"| __truncated__ "Mopani Copper Mines has declared a Force Majeura citing the effects of Covid 19 and the drop of copper prices w"| __truncated__ ...
 $ Sentiment    : chr  "Extremely Negative" "Neutral" "Extremely Positive" "Extremely Negative" ...
```

Table 5. Details of target feature

```
Factor w/ 5 levels "Extremely Negative",..: 1 4 2 1 2 1 3 1 2 5 ...
Extremely Negative: 178
Extremely Positive: 241
Negative: 330
Neutral: 271
Positive: 408
```

5.3 Text Preprocessing

A string of text made up of special characters, spaces, words, and numbers can be found under the "OriginalTweet" feature. The first step in text preprocessing is to create a corpus, or collection of text documents. 2. Making text all lowercase 3. Eliminating punctuation, numbers, filler words (to, and, but, or), blank or white spaces, 4. Divide the text messages into separate tokens. A sparse matrix contains a representation of the tokens. The number of words in each phrase is stored in each cell of the matrix, and the document's length is likewise decreased. Below is the outcome of the first three documents, both before and after preprocessing:

Table 6. Before and after the text preprocessing

Before Preprocessing	After Preprocessing
$\`1\` 'It may take 9 months for the economy to recover from the 19 shock says Bank Chairman Deepak Parekh He warns prices may fall by at least 20 leaving many companies bankrupt Banks too may be on a tighter leash post the' $\`2\` 'So Brookside buys 1litre of milk from farmers at about 35 removes fat to make ghee butter and gives you back water of 1litre Same happens for maize and wheat Kenyans so long as your leaders are in business with us we will always get conned' $\`3\` 'To health workers To security officers To artists To teachers To scientists To grocery store workers To food producers To volunteers To parents To YOU Thank you for doing your best to help fighting Together we will defeat the '	$\`1\` 'may take month economi recov shock say bank chairman deepak parekh warn price may fall least leav mani compani bankrupt bank may tighter leash post' $\`2\` 'brooksid buy litr milk farmer remov fat make ghee butter give back water litr happen maiz wheat kenyan long leader busi us will alway get con' $\`3\` 'health worker secur offic artist teacher scientist groceri store worker food produc volunt parent thank best help fight togeth will defeat'

The following figure and table display the top 15 words used most frequently:

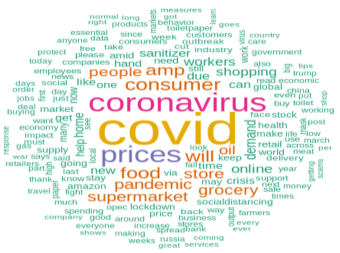

Fig. 1. Frequently Used Words

The content of document term matrix with three rows is shown in the following table:

Table 7. Three rows of Document Term Matrix

1. 'DocumentTermMatrix'
2. 'simple_triplet_matrix'
3. <<DocumentTermMatrix (documents: 3, terms: 10)>>
   Non-/sparse entries: 0/30

   Sparsity : 100%

   Maximal term length: 8

   Weighting : term frequency (tf)

   Sample :

   Terms

```
Docs abbott abl absolut abus acceler accept access accommod accord account
   1      0   0       0    0       0      0      0        0      0       0
   2      0   0       0    0       0      0      0        0      0       0
   3      0   0       0    0       0      0      0        0      0       0
```

Creating a data frame from the document matrix so that a naïve Bayes classifier may be constructed is the last stage in the data preparation process. Usually, categorical features in the data are used to train the naïve Bayes classifier. The 1 and 0 values are converted to factor with labels No and Yes using the factor command. Lastly, the training and test data that has been recently altered and updated to include 1428x2197 is displayed below:

Table 8. First six rows of the final dataframe

A data.frame: 6 × 2197																					
	abbott	abl	absolut	abus	acceler	accept	access	accommod	accord	account	⋯	youcv	young	your	yourcustomersaythankyou	youtub	youv	zafar	zero	zonec	Class
1	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	⋯	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**1**
2	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	⋯	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**4**
3	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	⋯	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**2**
4	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	⋯	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**1**
5	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	⋯	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**2**
6	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	⋯	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**1**

5.4 Model Building

The raw text messages from twitter have been turned into a data frame that may be represented by a statistical model. The naïve Bayes model has been selected for sentiment categorization. This classifier has shown to produce better results and is widely used in natural language processing. Based on the "Bayes" theorem, it calculates the likelihood that a given text in a tweet is extremely negative, extremely positive, negative, neutral, or positive.

5.4.1 Split Dataset and Training the model

For training and testing, the data set was divided into 75:25 ratios, respectively. The following table provides the same's size:

Table 9. Split of dataset

Subset Taken[based on a specific date]	1428.2
Corpus Content	1428 documents
Document Term Matrix	1428.2197
Final Train Final Test	1058.2197 370.2197

After preprocessing and split of dataset into train and test, naïve Bayes algorithm has applied to estimate the class of sentiments in the text document. Predictions have been made using the classifier object's output.

5.4.2 Train the model and Prediction

The naïve Bayes machine learning model received the train control function as inputs along with the parameters "repeatedcv" (resampling technique), 10 folds, 2 counts for repeats, and the laplace estimator set to 1. Below is the estimated system time:

user system elapsed 0.885 0.000 0.899

5.4.3 Evaluating Performance of the model

The trained classifier was assessed by passing the test data, the trained model, and the type "class" as arguments to the predict function. The sample of non-equal samples that was computed is displayed in the following table:

Table 10. 41-Sentiments are misclassified

A data.frame: 6 x 2		
	Predicted_Sentiment Level	Actual_Sentiment Level
1	4	4
2	2	2
3	3	3
4	2	2
5	2	2
6	4	4

5.4.4 Other measures of performance

We must test the classifier object's predictions on the test data's unseen text in order to assess it. After a classifier was trained and the trained model was assessed using the test data set, a cross table and confusion matrix were produced. The following tables display the results' cross table and confusion matrix:

Chi-Square Contribution **Total Observations in Table: 370**						
final$Actual	**final$Predicted** **1**	**2**	**3**	**4**	**5**	**Row Total**
1	51 115.790 1.000 0.554 0.138	0 7.305 0.000 0.000 0.000	0 10.476 0.000 0.000 0.000	0 10.200 0.000 0.000 0.000	0 10.338 0.000 0.000 0.000	51 0.138
2	0 13.178 0.000 0.000 0.000	53 271.592 1.000 1.00	0 10.886 0.000 0.00	0 10.600 0.000 0.00	0 10.743 0.000 0.00	53 0.143

		0 0.143	0 0.000	0 0.000	0 0.000	
3	17 1.622 0.183 0.185 0.046	0 13.322 0.000 0.000 0.000	76 169.468 0.817 1.000 0.205	0 18.600 0.000 0.000 0.000	0 18.851 0.000 0.000 0.000	93 0.251
4	1 16.702 0.013 0.011 0.003	0 10.743 0.000 0.000 0.000	0 15.405 0.000 0.000 0.000	74 232.067 0.987 1.000 0.200	0 15.203 0.000 0.000 0.000	75 0.203
5	23 0.077 0.235 0.250 0.062	0 14.038 0.000 0.000 0.000	0 20.130 0.000 0.000 0.000	0 19.600 0.000 0.000 0.000	75 153.028 0.765 1.000 0.203	98
Column Total	92 0.249	53 0.143	76 0.205	74 0.200	75 0.203	**370**

Table 11. Cross Table

Confusion Matrix and Statistics					
	Reference				
Prediction	**1**	**2**	**3**	**4**	**5**
1	51	0	17	1	23
2	0	53	0	0	0
3	0	0	76	0	0
4	0	0	0	74	0
5	0	0	0	0	75

Overall Statistics

Accuracy	0.8892
95% CI	(0.8527, 0.9193)
No Information Rate	0.2649
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.8614
Mcnemar's Test P-Value	NA

Statistics by Class:

	Class: 1	**Class: 2**	**Class: 3**	**Class: 4**	**Class: 5**
Sensitivity	1.0000	1.0000	0.8172	0.9867	0.7653
Specificity	0.8715	1.0000	1.0000	1.0000	1.0000
Pos Pred Value	0.5543	1.0000	1.0000	1.0000	1.0000
Neg Pred Value	1.0000	1.0000	0.9422	0.9966	0.9220
Prevalence	0.1378	0.1432	0.2514	0.2027	0.2649
Detection Rate	0.1378	0.1432	0.2054	0.2000	0.2027
Detection Prevalence	0.2486	0.1432	0.2054	0.2000	0.2027
Balanced Accuracy	0.9357	1.0000	0.9086	0.9933	0.8827

Table 12. Confusion Matrix for the improved model

6. Analysis of Review

This article examines a subset of data, specifically on date that includes five levels of sentiment: extremely negative, extremely positive, negative, neutral, and positive. This is done as part of our search effort. The confusion matrix is a more thorough method of evaluation that gives researchers further insight into the performance of their models. It is used to assess the performance of machine learning models. The diagonal table items in the confusion matrix above indicate which feelings from the tweets should be correctly classified. In other words, out of the 76 samples, 74 were neutral, 75 were positive, 51 were extremely negative, 53 were extremely positive, and 76 were negative.

The model now classifies 23 samples that were predicted to be in the positive class as highly negative, 1 sample that was expected to be in the neutral class as well as extremely negative and 17 samples that were expected to be in the negative class as extremely negative. The zero value indicates that the classifier did a good job of learning the data, meaning the model does not misidentify samples that were initially part of the respective classes. The kappa value was 86% and the total accuracy was 89%. The classification model also demonstrated very strong agreement for the other metrics, including sensitivity and specificity.

Tables of probabilities that are helpful in estimating the likelihood of the different classes are produced by the naïve Bayes algorithm. The Bayes theorem was used to calculate the relationship between the sentiment levels. This technique works best with huge text-based datasets and useful for filtering texts according to their content levels. R packages for text preprocessing, building models and visualization include tm, wordcloud, e1071, caret, and gmodels. Almost 89% of the text could be classified by the algorithm using levels like extremely negative, extremely positive, neutral, positive, and negative sentiments.

7. Conclusion

We can access a vast amount of data from social media with just one click. Sentiment analysis can be used to analyze data from social media platforms and anticipate user sentiments, including extreme negativity, extreme positivity, positivity, negativity, and neutrality. By examining social media information, sentiment analysis aims to enhance people's mindsets, levels of sales, service, or product quality. It also helps to establish new strategies. This can be accomplished by devising a novel method for selecting features or subsets from the dataset.

Predicting the invisible ideas that go through a person's head is difficult. A person's mental capacity varies depending on their gender, the issue at hand, the location, and the frequency of their experiences. Offering advice on how to solve an issue can help someone get out of a difficult situation. In order to do this, the researcher can forecast the participants' thoughts by tracking their text, image, audio, and video discussions over time in social media networks. Any discipline can use sentiment analysis with supervised or unsupervised machine learning methods. This research study aims to improve and comprehend the sentiment analysis process by using the naïve Bayes classification algorithm on a subset of records from the entire dataset.

References

[1] Divya Bharathi G, Dr.Jagan A, Pradeep Kumar V, (2020). Toxic Sentiment Identification using R Programming., *International Journal of Engineering Technology and Management Sciences*[IJETMS].

[2] Zafarani, R., Abbasi, M.A., Liu, H, (2014). Social Media Mining: An Introduction.Cambridge University Press. Cambridge University Press, ISBN 978-1-107-01885-3.

[3] DoaaMohey El-Din Mohamed Hussein, (2016). A survey on sentiment analysis challenges. *Journal of King Saud University – Engineering Sciences.*

[4] Arielle Hesse, Leland Glenna, Clare Hinrichs, Robert Chiles, Carolyn Sachs, (2019). Qualitative Research Ethics in the Big Data Era. American Behavioral Scientist.

[5] Casey Fiesler, Nicholas Proferes., (2018). Participant Perceptions of Twitter Research Ethics. Social Media + Society.

[6] Alexandra Balahur, (2013). Sentiment Analysis in Social Media Texts. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics*, Atlanta, Georgia, pages 120–128.

[7] Tim O'keefe, Irena Koprinska.,(2009). Feature Selection and Weighting Methods in Sentiment Analysis. *Proceedings of the 14th Australasian Document Computing Symposium, Sydney,* Australia.

[8] Erik Tromp, Mykolapechenizkiy, Mohamed Medhat Gaber., (2017). Expressive modeling for trusted big data analytics: techniques and applications in sentiment analysis. Tromp et al. Big Data Analytics.

[9] Michael Zimmer., (2018). Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity. Social Media + Society.

[10] Brett Lantz, (2013). Machine Learning with R. Packt Publishing Ltd, ISBN 978-1-78216-214-8.

[11] Widiastuti, N. I.,(2018). Deep Learning – Now and Next in Text Mining and Natural Language Processing. INCITEST, IOP Conf. Series: Materials Science and Engineering.

[12] Sahar Sohangir, Dingding Wang, Anna Pomeranets And Taghi, M. Khoshgoftaar.,(2018). Big Data: Deep Learning for financial sentiment analysis. Sohangir et al. J Big Data.

[13] Ha Sung Hwang., (2019). Why Social Comparison on Instagram Matters: Its impact on Depression. KSII Transactions on Internet and Information Systems.

[14] Concetta Papapicco., (2019). WhatsApp at Work? Instant Messaging Improves Professional Well-Being. SRG *International Journal of Communication and Media Science* (SSRG-IJCMS) - Volume 6 Issue 1.

[15] Chinthapanti Bharath Sai Reddy, Kowshik, S., Rakesh Kumar, M.V., O.Nikhil Kumar Reddy, Gopichand, G., (2020). Analysing and Predicting the Emotion of WhatsApp Chats Using Sentiment Analysis. ISSN: 0193-4120.

[16] Abubakar Ahmad, Nuhu Abdulhafiz, A. And Abubakar Abdulkadir. (2021). A Comprehensive Data Analysis on FUDMA ASUU Whatsapp Group Chat. FUDMA *Journal of Sciences (FJS*), ISSN online: 2616-1370, ISSN print: 2645 – 2944, Vol. 5 No. 2, pp. 26 – 33.

[17] Saikat Dutt, Subramanian Chandramouli, Amit Kumar Das, (2019). Machine Learning. Pearson India Education Services Pvt.Ltd, ISBN 978-93-530-6669-7.

[18] ReferenceLink- https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification.