



A Comprehensive Analysis of Approaches for Sentiment Analysis Using Twitter Data on COVID-19 Vaccines

Amrita Mishra¹, Mohd. Saif Wajid², Upasana Dugal³

^{1,2,3}Department of Computer Science Engineering, BBD University, Lucknow, India
mishra17amrita@gmail.com¹, mohdsaif06@gmail.com², upasana_gupta31@bbdu.ac.in³

How to cite this paper: A. Mishra, Mohd. S. Wajid, U. Dugal (2021) A Comprehensive Analysis of Approaches for Sentiment Analysis Using Twitter Data on COVID-19 Vaccines. *Journal of Informatics Electrical and Electronics Engineering*, Vol. 02, Iss. 02, S. No. 009, pp. 1-10, 2021.

<https://doi.org/10.54060/JIEEE/002.02.009>

Received: 05/04/2021

Accepted: 25/05/2021

Published: 05/06/2021

Copyright © 2021 The Author(s).

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0>



Open Access

Abstract

Sentiment Analysis has paved routes for opinion analysis of masses over unrestricted territorial limits. With the advent and growth of social media like Twitter, Facebook, WhatsApp, Snapchat in today's world, stakeholders and the public often takes to expressing their opinion on them and drawing conclusions. While these social media data are extremely informative and well connected, the major challenge lies in incorporating efficient Text Classification strategies which not only overcomes the unstructured and humongous nature of data but also generates correct polarity of opinions (i.e. positive, negative, and neutral) . This paper is a thorough effort to provide a brief study about various approaches to SA including Machine Learning, Lexicon Based, and Automatic Approaches. The paper also highlights the comparison of positive, negative, and neutral tweets of the Sputnik V, Moderna, and Covaxin vaccines used for preventive and emergency use of COVID-19 disease.

Keywords

Sentiment Analysis (SA), Machine Learning (ML), Supervised Learning, Unsupervised learning, Twitter

1. Introduction

Today the world is a Machine Dependency era. Well-formed systems for information exchange from peer to peer or B2B are established. The need of the hour is to ensure that besides navigating the data soil, customer's sentiments are evaluated. The correct assessment of user sentiments proves to be highlighting feature in winning or losing the product's name and growth in market. Earlier the information and feedback exchange systems were file and paper based which was accessible by limited people. However, today social media like Twitter serves as major platforms where users freely expresses their opin-



ions and it is accessible within remote areas. The developers can even analyze the tweets based upon selective geographic locations and form conclusions on regional basis.

Sentiment analysis (SA) is the area which deals with judgments, responses as well as feelings, which is generated from texts, being extensively used in fields like data mining, web mining, and social media analytics because sentiments are the most essential characteristics to judge the human behavior [29]. Customer sentiments can be found in tweets, comments and reviews. For example, Reviews delivered by customer on online sites after purchasing the product, property or visiting the hotels. Sentiment Analysis plays a vital role in Data Science as it brings forward a computational study about the diverse opinions. The calculative study provides a platform to derive the true meaning of customer opinion whether positive or negative or neutral.

What more that sometimes even if customer gives bias comments like, “The A.C. Compressor is working well but costly one.” can be examined by machine to draw accurate conclusion. Furthermore, dwelling into study of Sentiment Analysis involves descriptive study of SA via Machine Learning and Lexicon based approaches. Some basic types of SA are Fine grained, Emotion based, Aspect based and Multilingual based. Fine grained is used when polarity precision is important for a business. Example, Very positive = 5 stars and Very Negative = 1 star. Emotion based aims at detecting emotions, like happiness, frustration, anger, sadness, and so on while Aspect Based analyze sentiments of texts, let’s say product reviews to know which particular aspects has positive, neutral, or negative way. Multilingual Analysis involves a lot of preprocessing and resources e.g. translated corpora or noise detection algorithms are the There has been a lot of past research on different strategies to use the web innovation to expand the advantages of clients and in addition organizations in the commercial center.

The worldwide spread of corona virus termed as COVID-19 by World Health Organization in 11 March, 2020 have challenged many International and National Research Institutes to discover useful vaccines. Russia developed Sputnik V on 12 August, 2020, U.S developed Moderna on 17 December, 2020, and India developed Covaxin on 03 January, 2021 for prevention and emergency use of COVID-19.

Twitter, one of the most popular micro blogging social networking site where people tweet their opinions in a concise manner, typically in less than or equal to 140 words [14]. Twitter platform is widely used to deliver tweets relating with vaccines. Hence Twitter datasets of vaccines have been used. The paper is arranged as follows: Section 2 is the related work of the studies conducted by researchers using Twitter Data. Section 3 describes the methodologies in machine learning and lexicon-based approaches for sentiment analysis. Section 4 consists of Table of Tools for sentiment analysis. Section 5 approaches towards results and discussions of the positive, negative and neutral sentiments for the vaccines among Twitter users in Pandemic.

2. Literature Survey

Sentiment Analysis has been of avid interest to researchers lately. A lot of work has been put into it and there is a vast domain of its applications. Gaurav Bhatt [21] has performed Sentiment Analysis over Educational institutions Using Twitter Dataset of IIT, NIT and AIIMS Colleges in India with SVM, Naïve Baye’s and ANN algorithms and accuracy of 89.6%.

The area of Neural Networks has been investigated for performing sentiment analysis on benchmark dataset consisting of online product reviews. Bessalov et al. [2] carried out binary classification on Amazon and Trip Advisor dataset using Perceptron classifier and obtained one of the lowest error rates among their experiments of 7.59 and 7.37 on the two datasets respectively. Researchers have also been working upon prediction of accuracy of tested datasets using Machine Learning Algorithms. Kanakraj and Guddeti [3] used Natural Language Processing techniques for Sentiment Analysis and compared Machine Learning Methods and Ensemble Methods to improve on the accuracy of classification. Shahheidari, et al. [4] used a



Naïve Baye's Classifier for classification and tested it for news, finance, job, movies and sportstaking into consideration Data Mining on basis of two emoticons(☺ and ☹).

Prediction of Election Results is another domain in which massive population expresses opinion over Social Networks. Rincy Jose and Varghese S Chooralil [7] have used Twitter Data with Classifier Ensemble Approaches with accuracy of 71.48% in predicting election results. Rincy, et al. [9] have also predicted election results with Word Sense Disambiguiton with accuracy of 78.6%.

Mohd Saif Wajid et al. [25] have used Sentiment Analysis Based on A.I Over Big Data. They have introduced the methodology for creating user recommended data group (Big data) by elaborating a matrix for user recommended data group for big data which is then reduced by dimension reduction technique.

Neethu M.S and Rajasree R [5] used twitter post on electronic products, compared the accuracy between different Machine Learning Algorithmn and further improved accuracy by replacing repeated character with two occurrences, including a slang dictionary and taking emoticons into consideration. Jotheeswaran and Koteeswaran(6) performed binary classification on the IMDB dataset by employing a Multi-Layer Perceptron Neural Network and using Decision Tree -Based Feature Ranking for feature extraction and a hybrid algorithmn(based on Differential Evolution and Genetic Algorithm)for weight training, thereby obtaining a maximum classification accuracy of 83.25%.

Laszlo and Attila (20) have used fresh scraped data collections over the Recurrent Neural Networks to determine what emotional manifestations occurred in given time interval in COVID-19. The Sentiment Analysis helps in monitoring area based upon the opinion raised in different territories.

3. Methodologies

3.1. Twitter

Twitter is a micro-blogging site where the user's posts comments and opinions related to services, products, activities, personalities in the form of tweets. Each user has a daily limit of 2,400 tweets and 140 characters per tweets [21]. Tweets of Sputnik V, Moderna and Covaxin are used as datasets. They are extracted using Twitter Developer account. The Credentials granted from Twitter Developer Account are connected with Python using Tweepy library. In this manner 2000 tweets for Sputnik V, Moderna and Covaxin have been extracted.

3.2. Machine Learning Approaches

They can be categorised in three fundament categories: Supervised, Unsupervised and Reinforcement Learning methods.

3.2.1. Supervised Learning

In this, with the input provided as labeled dataset, a model can learn from it. In labeled dataset the answer or solution to it is given as well. Major steps include, loading labeled input dataset, training model and testing. So, a labeled dataset of animal images would tell name of animal. It is further classified to Classification and Regression. The Classification algorithm predicts a discrete value that can identify the input data as a member of particular class or group. The Linear Classifier includes Support Vector Machine (SVM) and Neural Networks. Rule Based Classifier Predicts the result within well-defined set of rules. The Probabilistic Classifier are categorised into Bayesian Network, Maximum Entropy and Naïve based. Naïve Baye's is based upon Baye's Theorem and for handling Big Data Maximum entropy is applied. The Regression problems are responsible for continuous data for example, predicting the diabetes status of a patient given the blood pressure, sugar level, etc. Here, the input has to be sent to machine for predicting diabetes according to previous instances.

3.2.2. Unsupervised Learning

Here, no complete and clean labeled dataset is provided. It focuses on self-organized learning that helps find previously unknown pattern in dataset without pre-existing models. Different algorithms like K-means, Hierarchical, PCA, Spectral Clustering, DBSCAN clustering are used in unsupervised learning. For any input X and response variable Y , suppose $f(X) = Y$, in supervised learning there can be two goals 1. $f(X)$ closely approximates Y , 2. Predict values of Y given X . In unsupervised learning there is no response variable Y . The clusters within dataset are identified based on similarity. It is more useful and dataset is less expensive.

3.2.3. Reinforcement Learning

An agent interacts with its environment by performing actions and learning from errors or rewards. It follows Trial and Error as there is no predefined data and supervision.

3.3. Automatic Approaches

The automatic approach shown in Figure 1, involves feeding a classifier with text as an input and obtaining the polarity category that is positive, negative, or neutral. It involves following two phases:

- **The Training Phase:** In this phase, the original text is divided into TAG and TEXT part. The tag is fed as whole to machine learning algorithm while Text is passed through Feature Extractor. The Feature Vectors for text are generated. The Tag and Feature vector of Text worked upon by machine learning algorithm produces Classifiers.
- **The Prediction Phase:** In this phase, input Text that has to be predicted is passed through Feature Extractor. Feature extractor generates the feature vectors. The Feature Vectors are fed into the Classifiers and suitable category which is positive, negative, neutral for Tag is obtained.

Feature Extractor and Feature Vector: A Tag is the predetermined classification or category that a Text fall into. Feature Extraction technique involves conversion of Text into numerical representation in vector form. In Feature extractor, ML uses Bag of words as dictionary, where a vector is obtained by comparison and transformation. For example, if we have defined our dictionary to following words { vaccine, accomplishing, for ,the ,beneficial} and we wanted to vectorize the text “The vaccine for Covid-19 is accomplishing” and “It is beneficial for masses” we would have following representations of text (1,1,1,0,0,1) and (0,0,1,1,0) as feature vectors. Multiple feature vectors are fed into classifiers.

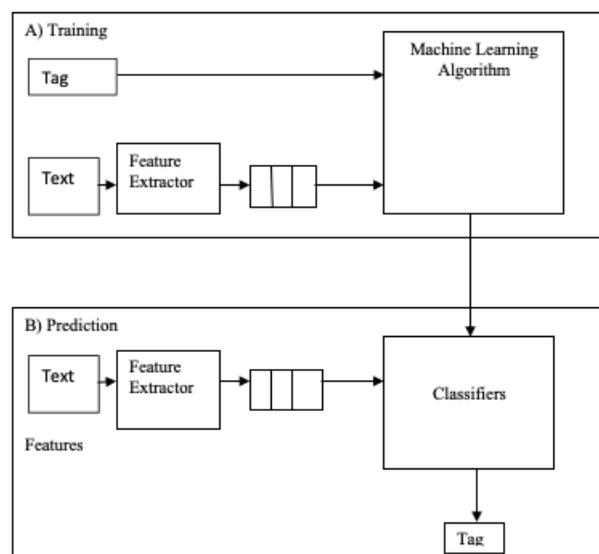


Figure 1: Automatic Approach Phases

3.4. Lexicon-based approaches

Dictionary put together methodologies for the most part depend with respect to a feeling vocabulary, i.e., a gathering of known and precompiled supposition terms, states and even figures of speech, produced for customary types of correspondence, for example, the SentiWordNet dictionary be that as it may, considerably progressively complex structures like ontologies, or lexicons estimating the semantic introduction of words or expressions can be utilized for this reason. Two sub characterizations can be found here: Dictionary-based and Corpus based methodologies.

3.4.1. Dictionary-based strategies

This involves the utilization of an underlying arrangement of terms (seeds) that are typically gathered and explained physically. This set develops via looking through the equivalent words and antonyms of a lexicon. A case of that lexicon may be WordNet, which was utilized to build up a thesaurus called SentiWordNet. The principle downside of this sort of methodologies is the lack of ability to manage space and setting explicit introductions; all things being equal, it may be an intriguing arrangement relying upon the issue.

3.4.2. The Corpus-based strategies

This emerged with the target of giving word references identified with an explicit area. These lexicons are created from a lot of seed sentiment terms that becomes through the pursuit of related words by methods for the utilization of either measurable or semantic systems. Regular Language Processing and Information Retrieval in Sentiment Analysis According to Cambria, Sentiment Analysis can be considered as an extremely limited NLP issue, where it is just important to comprehend the positive or negative estimations concerning each sentence as well as the objective elements or themes. In any case, regardless of being a limited issue, all works in this field, and all works in Information Retrieval, dependably battle with NLPs uncertain issues (invalidation taking care of, named element acknowledgment, word-sense disambiguation,) which are fundamental to recognize scholarly gadgets, for example, incongruity or mockery and thus, to discover and rate conclusions. The three dimensions of investigation that decides the distinctive undertakings of Sentiment Analysis are: (I) report level, (ii) sentence level and (iii) element/angle level. Report level thinks about that a record is an assessment on a substance or part of it. This dimension is related with the undertaking called report level opinion characterization. Notwithstanding, in the event that a report gives a few sentences managing distinctive viewpoints or elements, the sentence level is progressively appropriate. Sentence level is firmly identified with the assignment called subjectivity order, which recognizes sentences that express verifiable data from sentences that express emotional perspectives and sentiments Feature-based Opinion Mining and Opinion Summarization. A significant number of these papers pursue indistinguishable general procedures from other Information Retrieval works did previously, however supplanting a few factual or semantic factors for angles identified with assumptions. In this way, the principle distinction between these works is the element determination process. TextBlob, AFINN, VADER (Valence Aware Dictionary for Sentiment Reasoning) are used in Python for Lexicon Based Sentiment analysis.

4. Sentiment Analysis Tools

Sentiment Analysis tools are used in different fields such as politics, finance, business, etc. Sentiment analysis tools are given in Table2[10] and we also refer to Big Data analytics tools by considering a Comprehensive Survey on Big Data Analytics[11].

Table 1: SA Tools and Techniques

Sentiment Analysis tools	Techniques
SentiWordNet	Lexical dictionary and scores obtained by semi-machine learning approaches
LIWC	Dictionary and sentiment classified categories
EMOTICONS	Emoticons contained in the text
SenticNet	Natural Language Processing for inferring the polarity at semantic level
Happiness Index	Affective Norms for English Words (ANEW) and scores
AFINN	Affective Norms for English Words (ANEW) but additional attention on language used in microblogging platforms
PANAS-t	Eleven-sentiment psychometric scale
Sentiment40	API that allows classifying tweets to classify classes positive, negative, neutral
NRC	Huge collection of human-provided words with their emotional tags.
EWGA	Entropy –weighted genetic algorithm
FRN	Feature relation network considering syntactic n-gram relations
Scikit-learn	Machine learning and has useful tools for text vectorization
NLTK	Natural Language Processing Library for Python
SpaCy	Provides a strong set of low-level functions for Natural language Processing and support for training Text Classifiers.
TensorFlow	Developed by Google, provides a low-level set of tools to build and train neural networks
Keras	Works with Neural Networks, RNN and CNN in python
PyTorch	Deep Learning Framework

5. Results and Discussions

Based upon the datasets containing 2000 tweets each for three vaccines Sputnik V, Moderna and Covaxin and implemented using TextBlob, Lexicon Based Approach following results have been drawn. The TextBlob script gets the tweets as input and returns the text's polarity in terms of sentiment score. The sentiment score lies in the range of -1 to 1. Hence, the tweets are classified as 'Negative' if the score is less than 0, 'Neutral' if the score is equal to 0, 'Positive' if the score is greater than 0.

Table 2: Number of distribution of tweets

Vaccine	Positive Tweets	Negative Tweets	Neutral Tweets
Sputnik V	47	42	1911
Moderna	226	238	1536
Covaxin	177	67	1756

Table 3: Percentage of distribution of tweets

Vaccine	Positive (%)	Negative (%)	Neutral (%)
Sputnik V	2.35	2.1	95.55
Moderna	11.3	11.9	76.8
Covaxin	8.85	3.35	87.8



- [6]. J. Jotheeswaran and S. Kodeeswaran, "Decision Tree Based Feature Selection and Multilayer Perceptron for Sentiment Analysis," *Journal of Engineering and Applied Sciences*, vol. 10, no. 14, pp. 5883-5894, Jan 2015.
- [7]. R. Jose and V. S. Chooralil, "Prediction of election result by enhanced sentiment analysis on Twitter data using Word Sense Disambiguation," in *International Conference on Control Communication & Computing India (ICCC)*, Nov 2015.
- [8]. M. S. Wajid, S. Maurya, & R. Vaishya, "Sentence Similarity based Text Summarization using Clusters," *International Journal of Scientific & Engineering Research*, vol.4, no. 5, pp.1959-1966, May 2013.
- [9]. R. Jose and V. S. Chooralil, "Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach," in *International Conference on Data Mining and Advanced Computing (SAPIENCE)*, 2016.
- [10]. S. Rajalakshmi, S. Asha, and N. Pazhaniraja, "A comprehensive survey on sentiment analysis," in *Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)*, 2017.
- [11]. J. Vijayaraj, R. Saravanan, P. Victor Paul, et al, "A comprehensive survey on big data analytics tools," in *Online International Conference on Green Engineering and Technologies (IC-GET)*, vol.2, pp.32-40,2016.
- [12]. A. C. Pandey, S.R Seth, et al., "Sarcasm Detection of Amazon Alexa Sample Set," *Springer Nature Pte Ltd. Advances in Signal Processing and Communications*, 2019.
- [13]. D. U. R. Babu, "Sentiment analysis of reviews for E-shopping websites," *Int. j. eng. comput. sci.*, vol.6, no.1, pp.19965-19968, 2017.
- [14]. M. Khurana, A. Gulati, and S. Singh, "Sentiment analysis framework of twitter data using classification," in *Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Dec 2018.
- [15]. N. Majumder, S. Poria, H. Peng, et al., "Sentiment and sarcasm classification with multitask learning," *IEEE Intell. Syst.*, vol. 34, no. 3, pp. 38–43, 2019.
- [16]. S. Rajalakshmi, S. Asha, and N. Pazhaniraja, "A comprehensive survey on sentiment analysis," in *2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)*, 2017.
- [17]. F. H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme," *Decis. Support Syst.*, vol. 57, pp. 245–257, 2014.
- [18]. S. Porwal, G. Ostwal, A. Phadtare, et al., "Sarcasm detection using recurrent neural network," in *Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 746-748, 2018.
- [19]. M. Bouazizi and T. Ohtsuki, "A pattern-based approach for sarcasm detection on twitter," *IEEE Access*, vol. 4, pp. 5477–5488, 2016.
- [20]. L. Nemes and A. Kiss, "Social media sentiment analysis based on COVID-19," *J. Inf. Telecommun.*, vol. 5, no. 1, pp. 1–15, 2021.
- [21]. N. Mamgain, E. Mehta, A. Mittal, et al., "Sentiment analysis of top colleges in India using Twitter data," in *International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)*, pp. 525-530,2016.
- [22]. P. Nambisan, Z. Luo, A. Kapoor, et al. "Social media, big data, and public health informatics: Ruminating behavior of depression revealed through twitter," in *48th Hawaii International Conference on System Sciences*, pp. 2906-291, 2015.
- [23]. T. Tran and K. Lee, "Understanding citizen reactions and Ebola-related information propagation on social media," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 106–111, 2016.
- [24]. P. Song and T. Karako, "COVID-19: Real-time dissemination of scientific information to fight a public health emergency of international concern," *Biosci. Trends*, vol. 14, no. 1, pp. 1–2, Mar 2020.
- [25]. S. Kumar, A. K. Singh, P. Singh, et al, "Sentiment analysis based on A.i. over big data," in *Proceedings of the International Conference on Data Engineering and Communication Technology*, Singapore: Springer Singapore, pp. 641–649, 2017.
- [26]. K. Chakraborty, S. Bhatia, S. Bhattacharyya, et al., "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media," *Appl. Soft Comput.*, vol. 97, p. 106-754, 2020.
- [27]. M. M. Truşcă, "Efficiency of SVM classifier with Word2Vec and Doc2Vec models," *Proceedings of the International Conference on Applied Statistics*, vol. 1, no. 1, pp. 496–503, 2019.
- [28]. M. Bilgin and I. F. Senturk, "Sentiment analysis on Twitter data with semi-supervised Doc2Vec," in *International Conference on Computer Science and Engineering (UBMK)*, pp. 661–666, 2017.
- [29]. K. Chakraborty, S. Bhattacharyya, R. Bag, et al., "Sentiment analysis on a set of movie reviews using deep learning techniques," in *Social Network Analytics*, Elsevier, pp. 127–147, 2019.
- [30]. S. Latif, M. Usman, S. Manzoor, "Leveraging data science to combat COVID-19: A comprehensive review," *IEEE Trans. Artif. Intell.*, vol. 1, no. 1, pp. 85–103, 2020.
- [31]. S. Doğan, A. Betin-Can, and V. Garousi, "Web application testing: A systematic literature review," *J. Syst. Softw.*, vol. 91, pp. 174–201, 2014.



- [32]. V.R. Madaan, K.K. Bhatia, S. Bhatia, "Understanding the role of emotional intelligence in usage of social media," in 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 586–591, Jan 2020.
- [33]. B. A., Ninan, Sooraj, S. R. Ananthkrishnan et al., "Literature Review Web Application Designed for Schools and Colleges," Imperial Journal of Interdisciplinary Research (IJIR), vol.2, no.4, pp.350-356, 2016.
- [34]. D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 3642-3649, 2012.
- [35]. H. A. Shiddieqy, F. I. Hariadi, and T. Adiono, "Implementation of deep-learning based image classification on single board computer," in International Symposium on Electronics and Smart Devices (ISESD), pp. 133-137, Oct 2017.
- [36]. T. Molla, B. Khan, & P. Singh, "A comprehensive analysis of smart home energy management system optimization techniques," Journal of Autonomous Intelligence, vol.1, no.1, pp.15-21, 2018.
- [37]. P. Singhal, P. Singh and A. Vidyarthi, "Interpretation and localization of Thorax diseases using DCNN in Chest X-Ray," Journal of Informatics Electrical and Electronics Engineering, vol.1, no. 1, pp.1-7, 2020.
- [38]. M. Vinny, P. Singh, "Review on the Artificial Brain Technology: BlueBrain," Journal of Informatics Electrical and Electronics Engineering, vol.1, no.1, pp.1-11, 2020.
- [39]. A. Sahani, P. Singh and A. Kumar, "Introduction to Blockchain," Journal of Informatics Electrical and Electronics Engineering, vol.1, no.1, pp. 1-9, 2020.
- [40]. M. Misra, P. Singh, "Energy Optimization for Smart Housing Systems," Journal of Informatics Electrical and Electronics Engineering, vol.1, no.1, pp. 1-6, 2020.
- [41]. K. Chane, F.M. Gebru, B. Khan, "Short Term Load Forecasting of Distribution Feeder Using Artificial Neural Network Technique," Journal of Informatics Electrical and Electronics Engineering, vol.2, no.1, pp. 1-22, 2021.