ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Opinion Mining on Twitter Data Using Machine Learning

Mr. M. Kishore Babu[1], M.Tech (Ph.D.),AssistantProfessor, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
Vinay Rajolu[2], **Manjunath Popuri**[3],**Srinivas Peram**[4], **Niharika Marri**[5]
[2,3,4,5] UG Students, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
kishorebabu@vvit.net[1], 19BQ1A05I9@vvit.net[2], 19BQ1A05I2@vvit.net[3],
19BQ1A05H6@vvit.net[4], 19BQ1A05D2@vvit.net[5]

DOI:10.48047/IJFANS/V11/I12/175

ABSTRACT

Twitter is a social media platform, where we can express our opinions about various events happening worldwide. Researchers are concentrating on opinion mining due to the exponential rise of electronic media-based society and online social networks. A method for figuring out how individuals feel about an event is called opinion mining, often known as sentiment analysis. This is done by using various machine learning algorithms. With the help of these algorithms, we can do sentiment analysis without reading tweets manually. The outcome of these models could help government and industry experts socialize their policies and products. This study investigates how emoji characters are used on social media and how that affects opinion mining.We used three machine learning models to determine the polarity of tweets, such as whether they were positive, negative, or neutral. After building the models, this method combines voting-based models, and additional comparison metrics are determined for each model. We applied the proposed ensemble model to a few tweets to show the emoji's usefulness.

KEYWORDS: Machine Learning, Opinion Mining, Polarity, Sentiment Analysis, Simple Majority Voting, Twitter.

1. INTRODUCTION

The key factors influencing sentiment analysis sometimes referred to as opinion mining, are the preferences and viewpoints expressed on social media and microblogging websites. Social media sentiment research is being employed as a crucial data source to achieve a few objectives, including discovering consumer unhappiness or product flaws, projecting stock market values, and even predicting election results. The most significant aspect of the opinions posted by people on microblogging websites like Twitter may be their positive and negative feelings. Natural Language Processing (NLP) is used in Sentiment Analysis (SA) to extract opinions from user input and, in most instances, classify them as negative, natural, or positive. Organizations or enterprises may use these extracted views to help track products, improve services, or forecast occurrences. Text is the primary form of communication for social media users and users of microblogging services.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Twitter allows for a 140-character maximum for tweets. The benefit of the feature is that it forces users to express themselves succinctly by using a finite number of characters. In contrast to data from other social networking sites, Twitter data may therefore be more consistently standardized. Emoji's use on the Internet has increased in recent years. Ideograms and smiley's allowed users to express their emotions more easily through text in emails and internet pages. Emoji's like "Face_with_smile" have altered how we interact on social media and microblogging platforms. People typically use them when it is more challenging to explain them verbally. With just one Emoji, a text message may become more expressive.

When a city's name is presented alone, it has no emotive value.Yet the material can have sentimental significance if the user additionally used an Emoji with this name. The Emoji sign, for instance, has a cheerful look on its face. Might be used to express someone's positive impressions of the city. Nonetheless, combining some brand names with the angry face Emoji might suggest a dislike of the business. Adding an Emoji character may give a message of sympathy more depth. According to research published in the Social Neuroscience journal [2], people's emotional reactions to emoticons, a more condensed version of Emoji's, are like those elicited by real faces. Emotional expressions are now more important than previously believed, according to Dr. Owen Churches of Flinders University's psychology department. For the first time, the "Face with Tears of Delight"emoji was picked as the Oxford Dictionaries Word of the Year in November 2015. This occurred following the widespread use of these Emoji symbols online, particularly on social networks.

The objective of this sentiment analysis is to classify the text into the three different levels of granularity—sentence, document, and aspect [1]. The two basic methods for assessing sentiment on Twitter are machine learning methodology and lexicon-based approach. In this study, we use machine learning techniques to extract sentiment from tweets.

Both supervised learning and unsupervised learning are common categories for machine learning techniques. With the supervised learning technique, the classifier learns from labelled data to create a model that is then used to predict the target class in subsequent classification challenges. In contrast, in the unsupervised learning strategy, the classifier learns to distinguish the provided input text from unlabelled data.

2. LITERATURE SURVEY

Sentiment analysis models have been examined in several papers. Different areas use different methodologies and models. Nevertheless, only a handful of the research examined

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022**

the use of Emojis on social media. The initial Emoji lexicon was put into place by Novak et al. [3]. 1.6 million tweets in 13 different European languages were collected, and 83 human annotators assigned them to negative, neutral, or positive categories. They found that 4% of the tweets they collected contained emoticons. The 751 most popular Emoji characters were then ranked using the emotion score of the plain text.To depict the various classes, Emoji characters were utilised. In our work, we converted the emoji into their text format by using an emot module instead of giving scores to them. Elbagir et al. [4] removed the emoticons and corresponding emojis in the pre-processing phase, as we discussed above, we can understand how the emojis are affecting the text. To improve sentiment analysis, researchers have used several different tactics.

The most widely used techniques in this regard are those that rely on machine learning and lexicons. Because of their increased adaptability and precision, machine-learning techniques are becoming increasingly popular among researchers. It is standard procedure to apply supervised machine learning algorithms to improve the precision and effectiveness of sentiment classification or prediction analysis.

According to Singh et al. [5], applying various machine learning classifiers can help us improve the outcomes of sentiment analysis.. They carried out their experiments by using the WEKA software tool which is very much useful for the classification of text in the text. Apporv et al. [6] stated that feature engineering with a tree kernel provides the best results instead of one-way classification. The author defines two classification models in the paper: 2-way classification models and 3-way classification models.

In a two-way classification, the sentiment is categorised as either positive or negative, and in a three-way classification, as positive, negative, or neutral. According to the author, tree-based kernels produced the best accuracy and feature model. Many academics have used deep learning for tweet classification and image classification [7, 8]. A Tweets Classifier for US Airline Corporations Sentiments was proposed by Rustam et al. [9]. Pre-processing was requested for the dataset by the researcher.

The impact of several feature extraction methods on classification accuracy, including TF, TF-IDF, and word2vec, has been studied. Also, the usage of long short-term memory (LSTM) was examined using a particular dataset. To handle similar elections, the researcher advises using a Voting Classifier (VC) in the study. For determining results, the voting classifier must rely on spatial estimation (SE), stochastic gradient descent classifier (SGDC),

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

and a straightforward ensemble approach. Precision, accuracy, recall, and F1-score were used to assess a range of ML classifiers as working metrics.

The results show that the recommended VC is superior to one of the phase actors in terms of effectiveness.The experiment also showed that the effectiveness of the machine-learning algorithm increases when TF-IDF is employed as a feature input. According to Deepika et al. [10], good pattern recognition and model combining can boost the performance of models. The author also concludes that using feature representation approaches like TF and TF-IDF increases the model's accuracy. [16-24]

3. PROBLEM IDENTIFICATION

From our Literature Survey, we noticed that in the pre-processing phase emojis had been removed. Therefore, converted the emojis to text format using the emot module and the resultant text was pre-processed by NLP and pre-processed tweets are fed input to different machine learning algorithms like SVM (Support Vector Machine), Decision Tree, Random Forest, and Voting Classifier.

Adding class labels to unlabelled data is one of the problems in opinion mining. When using user reviews or comments as the source of the data, it is crucial to prepare a dataset for sentiment analysis by adding labels.

4. METHODOLOGY

This section describes the approach that was employed in this project work. Five key modules in the system are being proposed. Data collection is the first module, which involves gathering tweets for sentiment analysis; the second module involves translating emojis into text using a predefined library. This dataset is pre-processed in the third module to transform and refine tweets into a dataset that can be conveniently used for later analysis. Due to the lack of class labels in the collected dataset, the fourth module focuses on adding them. The essential features are then extracted to develop a classification model. The last module classifies tweets into three categories: positive, neutral, and negative using a variety of machine learning classifiers.

4.1 DATA COLLECTION

A dataset containing Tweets regarding NASA's most recent project, Artemis, is gathered from data.nasa.gov.in. It is an unlabelled dataset because it lacks any class labels. The dataset consists of 14110 tweets that are not classified as positive, neutral, or negative. One of the main responsibilities of data scientists or any other professional analyst is adding

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

labels to datasets, so adding labels to this dataset will be completed in the upcoming modules.

4.2 CONVERSION OF EMOJI'S

Emojis can provide crucial contextual information about the sentiment or emotion conveyed in a sentence, hence converting them to text is critical in sentiment analysis.Emojis can provide emotional indicators that sentiment analysis algorithms can better understand by being translated into English.

When a sentiment analysis system recognizes the positive emotion represented by the smiley face emoji, it will correctly classify a message with the phrase "I am so happy!" as having positive sentiment. The Python emot package is used to perform this conversion. The Python emot module is a small package that allows you to work with emojis and emoticons. Users can translate emoticons and emojis into matching Unicode characters or text equivalents.

4.3 PREPROCESSING OF TWEETS

Raw tweets are cluttered with noise and packed with slang and acronyms. Such noisy factors frequently affect how sentiment analysis methods perform. As a result, several preprocessing techniques are used before feature extraction. The following steps are involved in preprocessing tweets:

1) Tokenization and case conversion:

Tweets were converted to lowercase and divided into relevant tokens.

2) Removal of Stop words.

3) Removal of HTML tags and URLs (Uniform Resource Locator).

4) Removing punctuations and other special characters.

5) Lemmatization: Words are reduced into their base forms.

4.4 SCORING AND EXTRACTION

4.4.1 ADDING CLASS LABELS

In sentiment analysis, adding class labels to a dataset entail categorizing each piece of text in the dataset according to the sentiment it conveys. Using a three-way categorization system, which assigns a label of "positive," "negative," or "neutral" depending on the sentiment indicated in the text, is the most used method for categorizing sentiment. For adding class labels to this work, we used VADER (Valence Aware Dictionary and Sentiment Reasoner). Specifically created for social media text, VADER is a vocabulary and rule-based sentiment analysis tool. VADER is built on a sentiment lexicon that has 7,500 lexical elements, including phrases and words that are often used in social media. On a scale from

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

-4 (very unfavourable) to +4 (highly positive), with 0 denoting neutrality, each lexical property is scored.

4.4.2 FEATURE EXTRACTION

For use in creating a classification model, we extract elements or features during the feature extraction stage. The features that are extracted from datasets that comprise raw data in various formats, such as text, a series of symbols, and images, are in a format that can be directly supported by machine learning algorithms. As a result, we employ several approaches, including count vectorizer and term frequency-inverse document frequency (TF-IDF), to extract features from the tweets. In this study, the feature matrix reflecting the significance of phrases to the corpus in a text is extracted using CountVectorizer.

4.4.2.1 CountVectorizer:

A feature extraction method that is frequently employed in sentiment analysis is CountVectorizer. A set of text documents are transformed into a matrix of token counts using this particular bag-of-words paradigm. Each column in the matrix represents a distinct token in the corpus, and each row in the matrix represents a document. In sentiment analysis, CountVectorizer is often used to generate a feature matrix from a corpus of text, which can subsequently be used to train a machine learning algorithm to predict the sentiment represented in a new text. Because it is straightforward, effective, and capable of handling enormous amounts of text data, CountVectorizer is a well-liked option in sentiment research.

4.5 MACHINE LEARNING TECHNIQUES

There are two primary ways for obtaining sentiment analysis from text (tweets): lexicon-based methods and machine-learning methods [11–13].We employ machine learning in this study. For text classification and sentiment analysis, Twitter employs a range of machine learning approaches. Machine learning techniques train the system with some specified training data with predictable outputs, enabling working with new test data.Support Vector Classifier (SVC), Decision Trees (DTs), and Random Forest (RF) are some of the machine learning techniques are used to construct this study's machine learning classifier.

4.5.1 SUPPORT VECTOR CLASSIFIER

Machine learning applications for classification and regression employ Support Vector Machines (SVMs), a type of supervised learning technique. SVMs look for the ideal hyperplane when classifying data in a way that maximizes the distance or distance between the various classes of datapoints.The name of this method is Support Vector Classifier

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

(SVC). SVCs are well known for their propensity to avoid overfitting and to generalize well to new data, making them excellent for use with high-dimensional datasets.

4.5.2 DECISION TREE

Both classification and regression problems may be accomplished using a Decision Tree, a well-liked supervised learning technique in machine learning. It works by constructing a tree-like representation of choices and possible outcomes. The decision tree method will iteratively divide the data into subsets that are as similar as possible in terms of the sentiment labels using the input features. The algorithm will select the feature that best separates the data into pure subsets at each split. The process is repeated until a stopping requirement, such as a maximum depth or a minimum amount of data points in a leaf node, is reached. The sentiment of new text reviews can be predicted using the implemented classifier.

4.5.3RANDOM FOREST

For classification and regression applications, Random Forest is a popular supervised learning method in machine learning. It is an ensemble learning technique that integrates various decision trees to produce a model that is more reliable and accurate. The method used by Random Forest for sentiment analysis is to build several decision trees, each trained on a random subset of the data and input attributes. The predictions generated by each decision tree are then averaged out to merge the decision trees. The sentiment of fresh text reviews can then be predicted using the resulting model.

4.5.4 VOTING CLASSIFIER

To increase overall classification accuracy, a voting classifier integrates the predictions of various machine learning algorithms. To develop the voting classifier in this project, the SVM, Decision Tree, and Random Forest algorithms are trained on the same training dataset. Each algorithm forecasts the sentiment for a certain tweet, and the final categorization of the algorithm is determined by considering the majority voting. Soft voting and hard voting are the two types of voting. We consider hard voting in this study.

4.5.5 EVALUATION METRICS

Accuracy, precision, recall, and F1 score

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

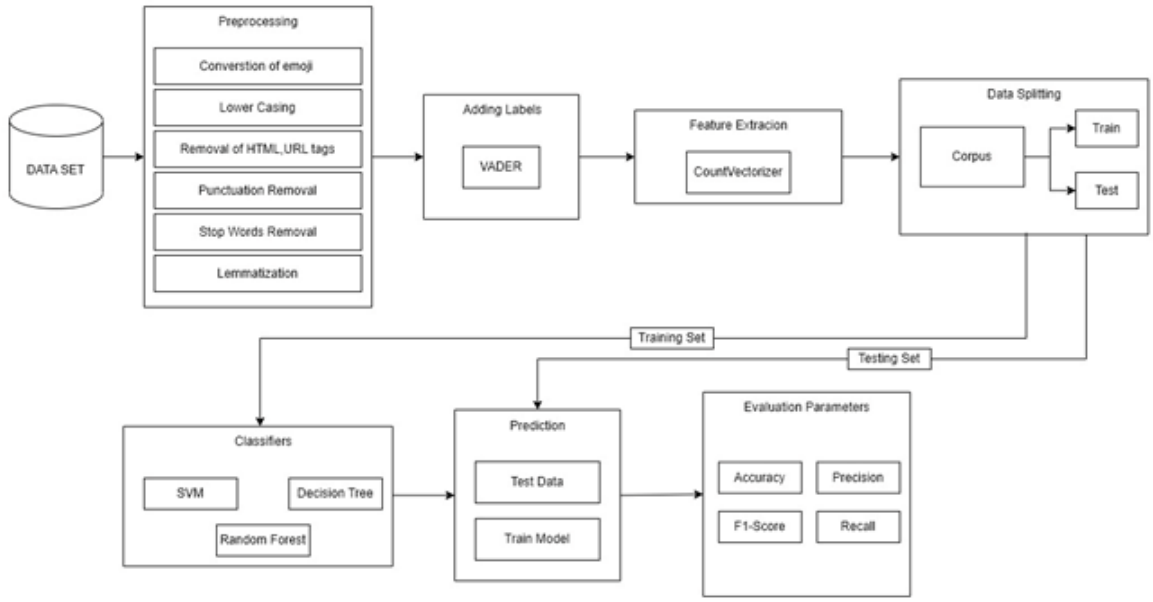

Figure 1. Proposed model architecture diagram

5. IMPLEMENTATION AND RESULTS

In the present study, we do Twitter opinion mining based on classification using three machine learning algorithms. The classification algorithms employed are SVM, RF, and DT (Decision Tree). Scikit-learn [14], [15], a Python-based open-source machine learning software tool, is used for the experiments. Many machine learning models are available in are often used assessment metrics for Categorizationissues.Accuracy is calculated as the fraction of cases out of all instances that are properly classified. Precision is the proportion of genuine positive predictions among all positive forecasts. Recall measures the ratio of true positives to all other positives. Precision and recall are both factors in the F1 score, which considers both false positives and false negatives. The Scikit-library for simple codeimplementation. We execute various tests utilizing machine learning algorithms aftertweet preprocessing, feature extraction, and data scoring. Here, a Voting classifier is an ensemble of Support Vector Classification(SVC), Decision Tree (DT), and Random Forest (RF) that gives the highest accuracy.

Table 1 shows the comparison of accuracy, precision, recall and f1-score of all four models.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Average Summary	Accuracy	Precision	Recall	F1-Score
SVM	86.1%	77%	78%	77%
Decision Tree	85.6%	76%	74%	75%
Random Forest	84.1%	85%	66%	68%
Voting Classifier	87.4%	80%	75%	77%

Table 1: Comparison of Models

6. CONCLUSION

This article describes the sentiment analysis of Twitter data for multi-class classification using a variety of machine learning techniques. In this article, we provide a technique for extracting sentiment analysis from Twitter and translating emojis into their equivalent text before sorting tweets into several ordinal categories using machine learning classifiers.Support vector classification, Decision Trees, and Random Forest are just a few of the classification approaches used in this work.

Experimental results indicate that Support Vector Classification has the highest accuracy than that Decision Tree and Random Forest. The voting classifier, with an accuracy of 87.40%, however, provides the best results. The experimental findings showed that the suggested model can detect emojis in text and convert them to their appropriate text before categorizing them. Also, if the dataset lacks any class labels, the model can produce various class labels for the text.

7. LIMITATIONS& FUTURE SCOPE

The considered dataset contains 14110 tweets, which are restricted in number, making classification challenging without additional information for each new context or tweet.The obtained data contains spelling errors and abbreviations that were not preprocessed and have an impact on the accuracy of machine learning models.

In the future, we intend to increase our model's accuracy by considering emojis not included in the library. Also, we would like to investigate additional deep learning and machine learning methods, including Deep Neural Networks, Convolutional Neural Networks, and Recurrent Neural Networks.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

8. REFERENCES

[1] R. Katarya, A. Yadav, "A comparative study of genetic algorithm in sentiment analysis", In 2018 2nd InternationalConference on Inventive Systems and Control (ICISC), Coimbatore, India, 2018, pp. 136-141.

[2] Churches, O., Nicholls, M., Thiessen, M., Kohler, M., &Keage, H. (Jan. 2014). Emoticons in mind: An event-related potential study. Social Neuroscience, 9(2), 196–202.

[3] Kralj Novak, P., Smailović, J., Sluban, B., &Mozetič, I. (Dec. 2015). Sentiment of emojis. PLOS One, 10(12).

[4] Elbagir, S., Yang, J. (Oct. 2019). Twitter Sentiment Analysis Based on Ordinal Regression.

[5] Jaspreet, Gurvinder, Rajinder. (2017). Optimization of sentiment analysis using machine learning classifiers. Doi:10.1186/s13673-017-0116-3

[6] Apoorv Agarwal, BoyiXie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau, "Sentiment Analysis of Twitter Data" Proceedings of the workshop on Language in social media (LSM 2011), 2011.

[7] M. Umer, S. Sadiq, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood, "A novel stacked CNN for malarial parasitedetection in thin blood smear images," IEEE Access, vol. 8, pp. 93782–93792, 2020.

[8] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, and B.-W. On, "Aggression detection through deep neuralmodel on Twitter," Future Gener. Comput. Syst., vol. 114, pp. 120–129, Jan. 2021.

[9] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. Choi, "Tweets classification on the base of sentiments for USairline companies," Entropy, vol. 21, no. 11, p. 1078, Nov. 2019.

[10] G. Deepika, and Dr.G. N.R. Prasad, "Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD)," Vol. 11, Issue 4, April 2022.

[11] X. Chen, M. Vorvoreanu and K. Madhavan, "Mining social media data for understanding students' learningexperiences", IEEE Trans. Learn. Technol., vol. 7, no. 3, pp. 246-259, Jul./Sep. 2014.

[12] N. R. Kasture and P. B. Bhilare, "An approach for sentiment analysis on social networking sites",Proc. Int. Conf. Comput. Commun. Control Autom., pp. 390-395, Feb. 2015.

[13] V. Singh and S. K. Dubey, "Opinion mining and analysis: A literature review",Proc. 5th Int. Conf.-Confluence Next Gener. Inf. Technol. Summit (Confluence), pp. 232-239, Sep. 2014.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved,** UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

[14] Scikit-Learn: Machine Learning in Python, Aug. 2019, [online] Available: http://scikit-learn.org/stable/.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikit-learn: Machine learning in Python", J. Mach. Learn. Res., vol. 12, pp. 2825-2830, Oct. 2011.

[16] Sri Hari Nallamala, et al., "A Literature Survey on Data Mining Approach to Effectively Handle Cancer Treatment", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 729 – 732, March 2018.

[17] Sri Hari Nallamala, et.al., "An Appraisal on Recurrent Pattern Analysis Algorithm from the Net Monitor Records", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 542 – 545, March 2018.

[18] Sri Hari Nallamala, et.al, "Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems", International Journal of Advanced Trends in Computer Science and Engineering, (IJATCSE), ISSN (ONLINE): 2278 – 3091, Vol. 8 No. 2, Page No: 259 – 264, March / April 2019.

[19] Sri Hari Nallamala, et.al, "Breast Cancer Detection using Machine Learning Way", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-2S3, Page No: 1402 – 1405, July 2019.

[20] Sri Hari Nallamala, et.al, "Pedagogy and Reduction of K-nn Algorithm for Filtering Samples in the Breast Cancer Treatment", International Journal of Scientific and Technology Research, (IJSTR), ISSN: 2277-8616, Vol. 8, Issue 11, Page No: 2168 – 2173, November 2019.

[21] Kolla Bhanu Prakash, Sri Hari Nallamala, et al., "Accurate Hand Gesture Recognition using CNN and RNN Approaches" International Journal of Advanced Trends in Computer Science and Engineering, 9(3), May – June 2020, 3216 – 3222.

[22] Sri Hari Nallamala, et al., "A Review on 'Applications, Early Successes & Challenges of Big Data in Modern Healthcare Management'", Vol.83, May - June 2020 ISSN: 0193-4120 Page No. 11117 – 11121.

[23] Nallamala, S.H., et al., "A Brief Analysis of Collaborative and Content Based Filtering Algorithms used in Recommender Systems", IOP Conference Series: Materials Science and Engineering, 2020, 981(2), 022008.

[24] Nallamala, S.H., Mishra, P., Koneru, S.V., "Breast cancer detection using machine learning approaches", International Journal of Recent Technology and Engineering, 2019, 7(5), pp. 478–481.