

Drug Recommendation System using Drug Reviews and Useful Count

Pavan¹, Dr. Vani K² and Rashmi Kulkarni³

¹⁻³Reva University, Bangalore, India

Email: pavanashrit@gmail.com, vanikrishna.s@reva.edu.in, kulkarni.rashmi.25@gmail.com

Abstract— In this era, we have recently seen the very dangerous spread of viruses all over the world called coronavirus. During this period we have seen there was a lack of hospital beds, doctors, healthcare workers, and lack of equipment. So many people died every day and night. The death rate was increasing like hell. Due to this situation, many people started to take medicine without any doctor consultation, but people did not know which medicine and how much dosage has to be consumed. That made the situation worsens. As we know computer technologies are growing rapidly, why can't we make use of these technologies to eradicate the worst situation like coronavirus spread? Yes, we can go for Machine Learning algorithms [3] which are having numerous applications. And so many automation works are running on machine learning algorithms [3]. This work is to develop a drug recommendation system [8] based on machine learning algorithms [3], that can help to reduce the specialist heap. In this work, we build the drug recommendation system by using the patient's reviews and useful counts of reviews to predict the sentiment using the various vectorization [2] methods such as Bow, TF-IDF, Word2Vec [2]. We build and compare the classification models such as ANN [8] and LSTM [6]. The best result is considered and used for the Drug Recommendation System. We recommend the top drug for a given disease. Finally, our work is going to be evaluated by precision, recall, flscore, accuracy score [2].

Index Terms— machine learning, vectorization, classification, ANN Algorithm, LSTM algorithm.

I. INTRODUCTION

The viral diseases are contagious. Viral infections are caused through different means. Virus that infects the skin are spread though the skin to skin contact. Likewise recently we have seen the Virus called novel coronavirus which caused the disease called COVID-19 in year 2019. The mutation of this virus lead to the Corona Virus2 (SARs-CoV-2). This caused the severe acute respiratory syndrome.

This virus mainly spreads though the means of air, where the infected person coughs, sneezes, or speaks. In less time many people gets infected and death rate are very high, all hospitals are occupied with COVID wards and many people have scarcity of beds in the hospital. And Many people start taking tablets without doctors prescription. That leads to many other side effects. Not only during covid situation but also in many villages, due to lack of medical facilities, people are tend to take tablets without doctors prescription. People may not be knowing which tablet is suitable for them. By seeing this all situation, we proposed this drug recommended system by using reviews of the patients.

In Era, the web based application are leading, if a person buys a Book, cloth or any materials, the feedback or what we call review is very important. If we visit any place also the Google asks us to give reviews on that place. These reviews or feedback helps the other user to buy the materials. The user will get the clear idea about the product and helps in happy shopping. Similarly why can't we have review for the drugs, for example if a person is having headache, if he gets any medicine for it. He can give recommendation, either positive or negative recommendation. Based on the reviews of many people, we are building the system called Drug Recommendation System.

In this proposed work, First we are collecting the data from repository, then cleaning and visualizing it. Next method is feature extraction, then followed by Training, sampling, classifying and then finally measuring the score by different metrics.

II. LITERATURE SURVEY

Here survey of some techniques used for telemedicine and drug recommendation system are discussed.

Youjun Bao et al. [1] designed and implemented a universal medicine recommender system framework that applies data mining technologies to the recommendation system [1]. Used machine learning algorithms such as SVM (Support Vector Machine) [11], BP neural network algorithm and ID3 decision tree algorithm based on the diagnosis data [1]. In this work they tried to apply these all three algorithms on the dataset and cross verified the result with help of accuracy and time taken. They found that with help of ID3 decision tree algorithm achieved 89% of accuracy in short time. But observed the bad scalability. Applied the BP neural network on the same dataset the accuracy was highest but the time taken to run is 1:7s [1]. So the third algorithm is SVM [11], when they tried with SVM it resulted with high accuracy of 95% and running time taken is 0.74seconds. So they choose the SVM model for the medicine recommendation system. It resulted with high accuracy, efficiency and scalability [1] for the open data set used in their work.

Satvik Garg et al. [2] proposed a drug recommendation system where predict the drug for the patients by using the patients reviews on the drugs. The sentimental analysis is done by using various machine learning algorithms and vectorization processes like Bow, TF-IDF, Word2Vec, and Manual feature analysis [2]. The classification algorithm are used and compared the result with metrics such as precision, recall, f1score, accuracy and AUC score [2]. Used Logistic regression, perceptron, Ridge Classifier, Multinomial NB, SGD Classifier and Linear SVC [2]. By comparing all the results by its evaluation, concluded that Linear SVC classifier with TF-IDF vectorization gives accuracy of 93%. So in this work they used SVC model for the drug recommendation system. Subhash chand Agrawal et al. [3] surveyed various machine learning algorithms for the sentimental analysis. They compared the accuracy of the models with different supervised and unsupervised algorithms. By the result analysis got to know that supervised machine learning algorithms are having the high accuracy over the unsupervised algorithms.

Prashant Upadhyaya et al. [4] proposed a model for the food items prediction using the sentimental analysis. In this work they collected the dataset in their own university food courts. The comparative analysis is done by using five machine learning algorithms such as Naïve Bayes, Naïve Bayes using TF-IDF techniques, Multi-Nomial NB, XG Boost, and Cat Boost [4] were selected. In this analysis, got to know that Cat-Boost algorithm gives the highest accuracy of 80.40% [4].

B. K. Bhavitha et al. [5] compared the machine learning algorithms for the sentimental analysis of the product reviews and the consumer attitude towards the newly launched product [5]. They compared the accuracy of each algorithms and their advantages and limitations. In result they declared that for the supervised machine learning models the accuracy score was 85% and it is higher than the unsupervised machine algorithms [5].

By observing this survey we got to know that for the sentimental analysis various machine learning algorithms are used for the creation of models and on top of it we can do classification by applying some specific classifiers. The model should be trained with positive and negative data and then tested. The accuracy of prediction is observed.

III. PROPOSED MODEL

The proposed model of drug recommendation system is having five different stages such as Data Preparation, Modelling, Evaluation, result comparison with different models and finally Drug Recommendation.

By using the collected Dataset, we are training the model in two different ways and then finally we are comparing the result of two models which gives high accuracy. According to the result comparison we are analysing the drug recommendatory system. Below fig 1 is the flowchart of the proposed model.

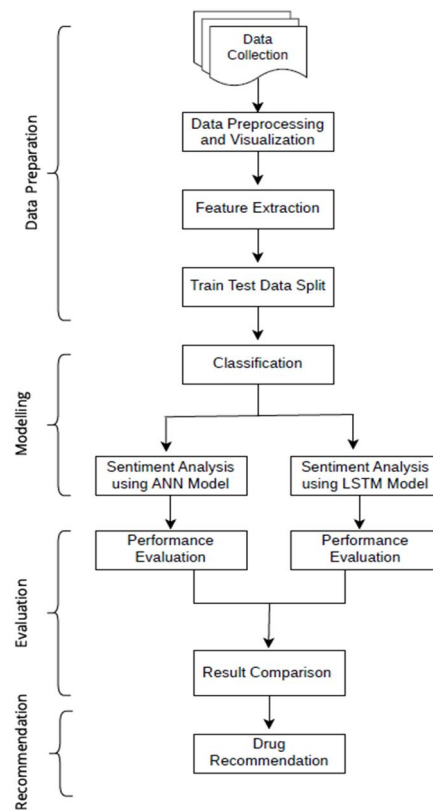


Figure. 1. Flowchart of Proposed Model

A. Data Collection

The Data collection is the first step of our proposed model. Collecting the meaningful data according to our needs is the big challenge. For this Drug recommendation system, the dataset is collected from the UCI ML repository. It is a Drug Review Dataset from the Drugs.com. This dataset contains 161297 of train data instances and 53766 of test data instances. The each dataset contains 7 attributes such as unique ID which is ID of unique patients review, second is Drug Name in text, third is condition in text, fourth is review in text, fifth is rating which is in range 1 to 10 where 10 is the best rating, sixth is date when review is given and last attribute is useful counts. Useful counts contain how many likes are for the particular review.

B. Data Pre-processing and Visualization

As our first step Data collection is done, we can't use the raw data directly for training and testing the models. We have to clean the data. By considering both training and testing data we have 215063 of instances in 7 columns. In data cleaning we are checking any attribute is having null values, if yes we are removing that particular instance from our data set.

The next step is Data visualization; in our work we are trying to visualize the data in 3 ways. In the first visualization we grouped the drug by its condition and plotted graph against drug counts and arranged them in descending order. We got to know that for the condition "pain" more drugs are available in our dataset. Here in the below fig 2, we plotted the graph for the top 40 conditions against drug counts.

In the visualization 2, we are working with review ratings given by users. In our dataset the ratings starts from 1 to 10. 10 is the best rating. So we are considering the ratings equal and above 5 is positive rating (binary 1) and ratings less than 5 is negative rating (binary 0). We added it as another attribute called "Review Sentiment" (column having the ones and zeros). We got 160613 instances as positive ratings and 53256 as negative ratings. We visualized it by plotting the pie chart in fig 3.1.

By seeing this we got to know that we are having more training instances than testing instances that should not be the case. So here considering only one condition which is having one drug we got 212053 of instances. Later taking the top 5000 instances of positive review and 5000 instances of negative review we made the exact equal amount of instances for both training and testing. It is visualized with figure 3.2.

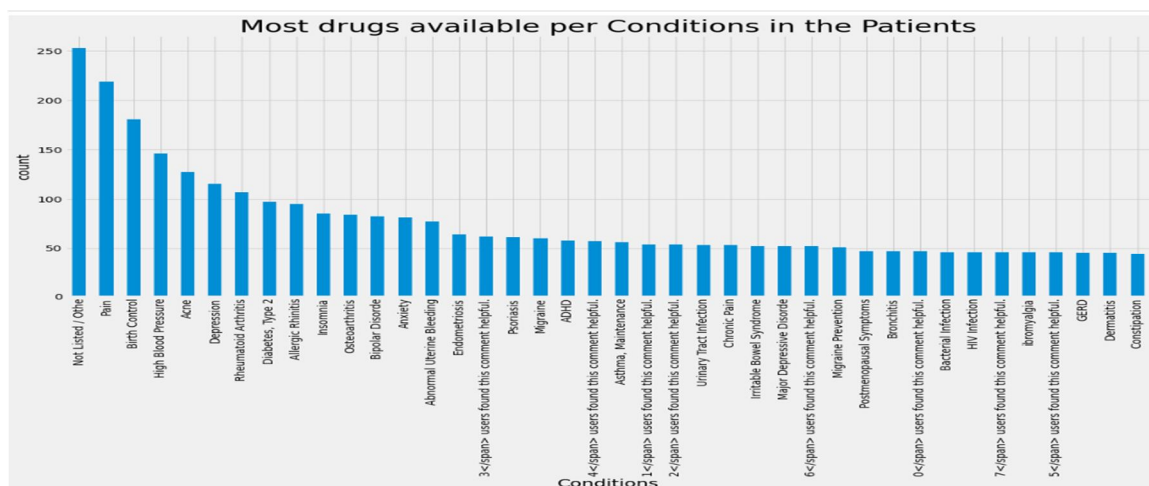


Figure. 2 Visualization-1

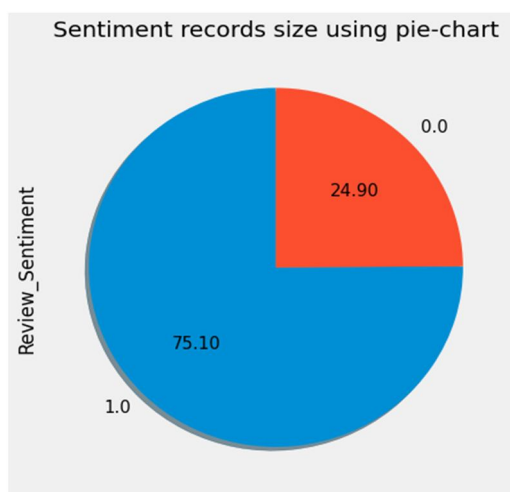


Figure.3.1 Visualization-2

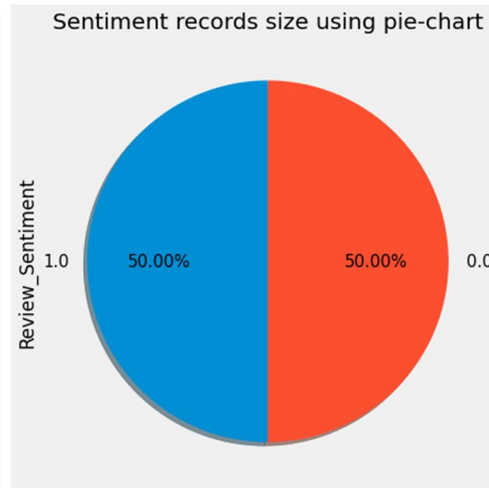


Figure. 3.2 Visualization-2

In the visualization 3, we tried to plot graph of year against the drug count of reviews in each year. We observed that in the recent years of data we are getting more reviews, this indicates the drug review is helping the people and its getting popular year by year. Here is the graph below fig 4.

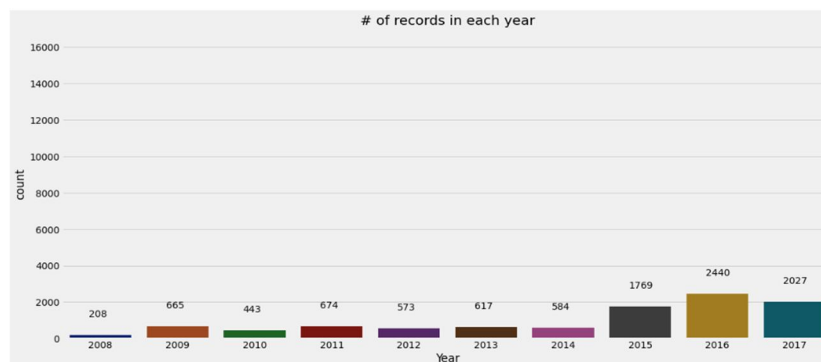


Figure.4 Visualization-3

C. Feature Extraction

As we are doing sentimental analysis [7][10] for the drug recommendation system. The preprocessing of data is required. Preprocessing of data means removing the unwanted piece of data from dataset and making it ready

for the sentimental analysis. In our proposed model we are extracting features by using machine learning algorithms like Bag of words (Bow) and vectorization using TF-ID Vectorizer and converting words to vector by fitting them in dimensions. We are using libraries like spacy, NLTK and WordCloud to get the frequency distribution of reviews. And using wordCloud and spacy to remove the stopwords. Lemmatization is done on the verbs. It not only removes the prefix and suffix but also guesses the better verb for it. Here is the result after preprocessing the review by lemmatization in fig 5.

uniqueID	drugName	condition	review	rating	date	usefulCount	Review_Sentiment	Year	month	day	processed_text	
0	88634	Prepopik	Bowel Preparation	"This was my 4th colonoscopy. For the first th...	1	2016-02-05	9	0.0	2016	2	5	colonoscopy use suprep not pleasant thing tast...
1	116276	Adalimumab	Psoriasis	"I am jealous of all those this drug has helpe...	1	2015-10-07	9	0.0	2015	10	7	jealous drug help develop issue dose doctor en...
2	101370	Aubra	Birth Control	"I love it!!! I hated taking birth control pil...	10	2016-02-20	13	1.0	2016	2	20	love hate birth control pill horrible effect a...
3	188589	ParaGard	Birth Control	"I had the paragard put in six weeks after my ...	1	2016-02-26	14	0.0	2016	2	26	paragard week son son big mistake life cramp u...
4	231649	Trazodone	Depression	"I take 50mg Trazodone before bed. It helps me ...	9	2015-04-15	104	1.0	2015	4	15	mg trazodone bed help sleep ease depression da...

Figure.5 Preprocessed Text after Data preprocessing

And the image we plotted with help of WorldCloud to get to know the most highly used words in large fonts and less used words in small sizes. Here is the result of our cleaned data set in fig 6.



Figure.6 WordCloud for Cleaned Data

D. Train Test Split

The Dataset should have the equal amount of instances for both train and test data, in our dataset we are having 75% of Train data and 25% of test data. So to make it equal we applied some functions and considered 50% of Train data and 50% of Test data for the sentimental analysis.

E. Classification

The classification is the technique use sentimental analysis for predicting the results. In our work we are training the model with positive and negative reviews, and then testing with test data to check the accuracy. In our work we have created two models, one with Artificial Neural Network using Machine Learning. And second by using LSTM algorithm.

Sentiment Analysis using ANN Model

This model is built by using the machine learning algorithm called ANN. We used tensorflow Keras library for sequential model. And Layers such as Dense, Activation and Dropout. Used the Adam algorithm for optimization of parameters. In Total we got 2,593,281. In that all are trainable. Created model with epoch of size 10 and batch size as 256. At the end we got validation loss as 16.5% . which reduced from 69.27%. Validation accuracy increased from 53.12% to 96.31%.

Sentiment Analysis using convolutional LSTM Model

In this model we used the another algorithm of machine learning that is LSTM(LongShortTermMemory). Here also we used sequential model from Keras library. The layers used in this model are more than CNN. The layers are Dense, Dropout, Activation, Conv1D, MaxPooling1D, Flatten, Embedding, LSTM, SpatialDropOut1D. Here we used tokenizer for the preprocessing of text. For convolution 1D algorithm used the activation layer of “relu”. And for the dense layer used the both “relu” and “sigmoid”. After applying Adam algorithm, we got total

parameters of 3,241,417, all are trainable. we used epochs of 10 and batch size as 16 to fit the model. Validation loss got decreased from the 69.31% to 4.95%. and Validation accuracy increased from 50.75% to 98.55%.

F. Comparing Models

By comparing the CNN Model and LSTM model we got to know that the Convolution LSTM is giving the high validation accuracy than the CNN. In training the model, Not only validation accuracy is considered but the reduction of validation loss is also very important. So by reviewing both model we got to know that by using the LSTM, we can reduce the validation more than the CNN. So we decided to use the LSTM model for the drug recommendation system.

G. Drug Recommendation System

In the Drug Recommendation system we are using the TextBlob library to check the polarity of the review sentiments. And used the Random Forest Classifier to predict the sentiment. We used the train model of LSTM and also useful counts from the dataset then Predicted the drug. If we give the condition, then it gives the prediction score for the drugs. We can consider these scores for the Drug Recommendation.

IV. RESULT AND ANALYSIS

In this proposed work we have done the analysis of the two classification models, such as classification model using Artificial Neural Network (ANN) and LSTM algorithm. In the result of ANN model the accuracy increased from 51.33% to 93.30% and Validation accuracy increased from 62.09% to 96.49%. Below is the graphical representation of accuracy with respect to epochs in fig 7. But the same test data applied with LSTM Model, the accuracy increased from 49.81% to 97.57% and the validation accuracy increased from 49.70% to 99.00%. The below graph represents the accuracy from the LSTM Model with respect to epoch in fig 8.

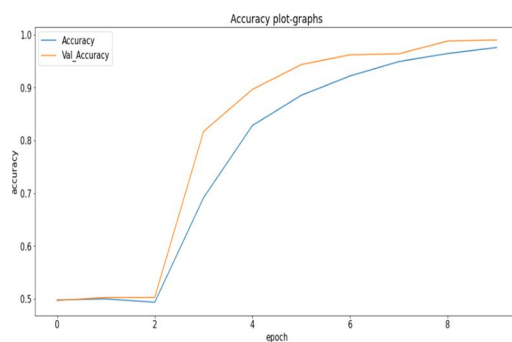


Figure .7 Accuracy Graph of ANN Model

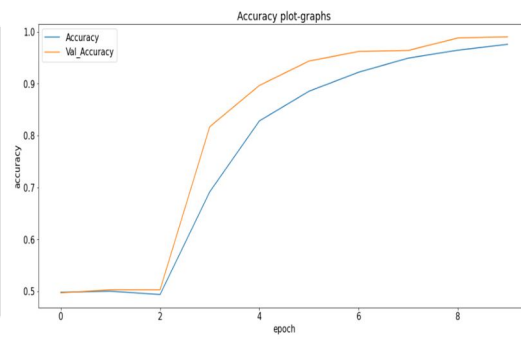


Figure.8 Accuracy graph of LSTM Model

By observing these two models result we can see that the validation accuracy is achieved more in the LSTM Model rather than ANN.

If we consider the loss percentage in both models, the ANN model the loss decreased from 69.30% to 23.17% and the validation loss decreased from 69.25% to 14.8%. Below fig 9 is the graph plotted for the loss in ANN model with respect to epoch. In LSTM model, the loss decreased from 69.32% to 7.72% and validation loss decreased from 69.33% to 4.62%. Below fig 10 is the graph plotted for the loss with respect to epoch.

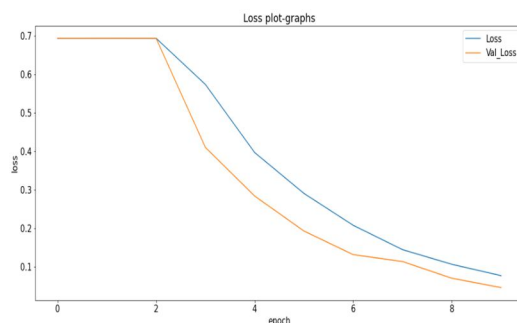


Figure.9 Loss graph of ANN Model

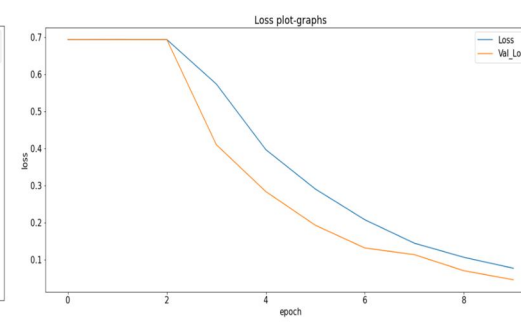


Fig.10 Loss graph of LSTM Model

As compared to ANN, the LSTM model is resulting the more validation accuracy. And also the reduction of loss is also very important in sentimental analysis. The validation loss decreased more by using LSTM model. So we are choosing here LSTM model for the drug recommendation system. Here are the precision, recall, f1-score and accuracy of both ANN and LSTM Models in table I.

TABLE I. RESULT ANALYSIS OF ANN AND LSTM MODELS

Model	Class	Precision	Recall	F1-score	Accuracy
ANN	Negative	0.96	0.97	0.97	0.96
	Positive	0.97	0.96	0.96	
LSTM	Negative	0.99	0.98	0.99	0.99
	Positive	0.98	0.99	0.99	

By applying Random Forest Classifier and the TextBlob library, we feed the dataset to the recommendation system on the LSTM model. If we give the input condition as Anxiety, the drug recommendation system gives recommendation score for the each drug of that condition. The user can use this recommendation score to consider the medicines for their needs. Here is the top 10 results shown below in fig 11.

Recommendation Score		
drugName	rating	
Tranxene	10	0.989791
Gabapentin	10	0.988041
Effexor XR	10	0.979517
BuSpar	10	0.971349
Atenolol	10	0.963949
Lexapro	10	0.958082
Pristiq	10	0.949691
Clonazepam	10	0.937948
Hydroxyzine	10	0.937854
Oxazepam	10	0.933883

Figure.11. Top 10 Recommendation for the Anxiety

V. CONCLUSION

In the proposed model, we trained and tested the data on two classification models that is ANN and LSTM [6], we found that the accuracy score is more in LSTM than ANN and also the reduction of validation loss is good in LSTM [6] than ANN. So we used the LSTM [8] model for the drug recommendation system. If we feed the condition for the drug recommendation system, it gives the recommendation score for the all the drugs available for that condition in the dataset. So in this system we didn't only make use of negative and positive reviews but also used the useful counts. Useful counts nothing but the likes for the reviews. So our Drug Recommendation System works with accuracy of 99% with LSTM [6] model and also by using useful counts it gives the recommendation scores.

FUTURE SCOPE

In our work we have worked only on the core part of our research, implementing this system with UI (User Interface) will be the future scope. UI may be web or mobile app development.

REFERENCES

- [1] Y. Bao and X. Jiang, "An intelligent medicine recommender system framework," 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), Hefei, China, 2016, pp. 1383-1388, doi: 10.1109/ICIEA.2016.7603801.
- [2] S. Garg, "Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 175-181, doi: 10.1109/Confluence51648.2021.9377188.

- [3] S. C. Agrawal, S. Singh and S. Gupta, "Evaluation of Machine Learning Techniques in Sentimental Analysis," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-5, doi: 10.1109/ISCON52037.2021.9702430.
- [4] P. Upadhyaya, S. Bhalla, A. Padhi, K. Rathi, A. Andey and R. Hooda, "Food Items Prediction Using Sentimental Analysis," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2022, pp. 751-756, doi: 10.1109/ICESC54411.2022.9885508.
- [5] B. K. Bhavitha, A. P. Rodrigues and N. N. Chiplunkar, "Comparative study of machine learning techniques in sentimental analysis," 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2017, pp. 216-221, doi: 10.1109/ICICCT.2017.7975191.
- [6] K. S. Vrunda and B. Jayasri, "Sentimental analysis of Twitter data and Comparison of covid 19 Cases trend Using Machine learning algorithms," 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2022, pp. 1-7, doi: 10.1109/GCAT55367.2022.9971980.
- [7] A. A. B. M, M. R, V. K and K. K. S R, "Sentimental Analysis for E-Commerce Website," 2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22), Nagpur, India, 2022, pp. 1-4, doi: 10.1109/ICETET-SIP-2254415.2022.9791606.
- [8] M. M. Rahman, M. M. Jahangir Kabir, M. Kabir and M. R. Karim, "Categorization of Product Feedback using a Deep CNN-based LSTM Model and a Comparative Study of Traditional ML Algorithms," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-6, doi: 10.1109/I2CT54291.2022.9825165.
- [9] M. D. Hossain, M. S. Azam, M. J. Ali and H. Sabit, "Drugs Rating Generation and Recommendation from Sentiment Analysis of Drug Reviews using Machine Learning," 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE), Bangladesh, 2020, pp. 1-6, doi: 10.1109/ETCCE51779.2020.9350868.
- [10] A. Vijayaraj, K. Bhavana, S. SreeDurga, S. Lokesh Naik, "Twitter based sentimental analysis of Covid-19 observations", Materials Today: Proceedings, Volume 64, Part 1, 2022 ,Pages 713-719,ISSN 2214-7853, <https://doi.org/10.1016/j.matpr.2022.05.194>.
- [11] P. Vijayaragavan, R. Ponnusamy, M. Aramudhan, "An optimal support vector machine based classification model for sentimental analysis of online product reviews", Future Generation Computer Systems, Volume 111, 2020, Pages 234-240, ISSN 0167-739X, <https://doi.org/10.1016/j.future.2020.04.046>.