Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Preprocessing of Drugs Reviews and Classification Techniques

Rosul Ibrahim Kazim [a]*, ***Enas Fadhil Abdullah*** [b]*

aDepartment of Computer Science, Collage of Education, University of Kufa, Najaf, Iraq. Email: russellalhusseini@gmail.com

bDepartment of Computer Science, Collage of Education for Girls, University of Kufa, Najaf, Iraq. Email: inasf.alturky@uokufa.edu.iq

ARTICLE INFO

Article history:
Received: 30 /05/2023
Rrevised form: 02 /07/2023
Accepted : 06 /07/2023
Available online: 30 /09/2023

Keywords:

Sentiment analysis, NLP, Machine Learning.

ABSTRACT

NLP refers to the computer-based study of language when referring to textual data such as texts or reviews. Natural language processing frequently seeks to give otherwise unstructured natural language a representation of the text that gives it structure. Marketing, social media, and customer relationship management are just a few of the businesses that heavily rely on sentiment analysis (SA), one of the methods used in opinion mining, typically assesses a textual review's structure to see if it conveys a positive or negative impression .In this paper, the challenge was to convert unstructured texts into structured texts for drug datasets, and interlocking between sentiment analysis, the mechanism that was adopted in the paper starts with several stages. The first stage is data preprocessing with natural language processing techniques, and the next stage is the prediction step with classification models are logistic regression(LR), random forest(RF), Naïve Bays(NB), and Support Vector Machine (SVM), the best Result of prediction accuracy 92%,for random forest .

MSC..

https://doi.org/10.29304/jqcm.2023.15.3.1261

1. Introduction

Natural Language preprocessing (NLP) is a theoretically supported collection of computer approaches for analyzing and modeling naturally occurring texts at one or more levels of linguistic analysis to achieve human-like

language processing for a range of activities or applications. NLP is an extremely efficient method for processing human language.

Sentiment analysis is a method for categorizing people's thoughts stated in reviews or surveys as positive, negative, or neutral using computational methods. Online evaluations now frequently include slang, emoticons, and other everyday language to better capture readers' opinions [1]. The Web has seen a huge flow of information, and it is still growing tremendously while giving users and customers access to a variety of resources about services like goods, hotels, and restaurants. Despite the benefits of such data, the overwhelming volume of alternatives and the huge flow of information present challenges for users [2] [3]. Our work in this paper exploits the concepts of natural

∗Corresponding author

Email addresses:

Communicated by ‘sub etitor’

language processors on the medical review to become comprehensible and to extract sentiment features and classification.

The remainder of the essay is structured as follows Section 2 represent Literature Reviews section ,3 ,4 and 5 represent concept NLP, sentiment analysis ,and machine learning respectively, section 6 concept Classification Techniques ,and section 7 Accuracy Measures Finally section 8 Methodology that used.

2. Literature Reviews

One of the newest sub-disciplines of natural language processing is sentiment analysis, which is the study of opinions on a certain topic from plain text. Drug sentiment analysis has grown significantly in recent years because categorizing medications according to their effectiveness by examining user feedback can help prospective customers learn more and make wiser decisions about a given drug Using sentiment analysis.

Table 1 - Literature Reviews

Reference number	Data set	Methodology	Classification algorithm and accuracy
[4]	UCI machine learning repository, drugs reviews	Utilized the "DRUG REVIEW" dataset's VADER emotional analysis to obtain structured data, and then utilized the available ML Algorithms.	Linear SVC 90.19 Naïve Bayes 79.57 Logistic Regression 85.27 SVM 89.34 Random forest 81.83
[5]	UCI machine learning repository, drugs, reviews.	A recommendation system based on a hybrid RNN stacked with bi-directional LSTM model.	Multinomial Naive Bayes 0.75354 Random Forest 0.829 Linear SVC 0.581 Logistic Regression 0.637 RNN-BiLSTM 0.839
[6]	UCI machine learning repository, drugs, reviews.	Applied tokenization and lemmatization on review.	Random Forest 94.06 Multilayer Perceptron 86.82 SVM 88.63 Naïve Bayes 88.57
[7]	UCI machine learning repository, drugs, reviews.	The following deductions were made using deep learning algorithms, which revealed a general trend where the Count Vector outperforms the TF_IDF encodings.	SVM 79.04 Logistic Regression 76.5 Random forest 70.2
[8]	UCI machine learning repository, drugs, reviews.	Opinion Model focuses on predicting the degree of drug satisfaction among patients.	Logistic 72 Regression 68.2SVM RL+ SVM 74.8
[9]	UCI machine learning repository, drugs , reviews.	built various classification models to classify textual reviews of medications accompanied by user ratings With drug reviews using TF-IDF features as input, numerous supervised machine learning classifiers were created, including Random Forest and Naive Bayesian classifiers.	BERT 84 ALBERT 78 Random Forest 68 Naïve Bayes 61 ELECTRA 86

3. Natural Language Processing (NLP)

The computational analysis of linguistic data is known as natural language processing (NLP), and it is most typically done on textual data such as papers or publications. NLP frequently seeks to produce text representations that give the unstructured natural language more structure, Natural language processing uses computer techniques to acquire, comprehend, and produce human language-based content. The development of important technologies like machine translation, speech recognition, and speech synthesis was the primary emphasis of early computational methods for language studies, As well as spoken dialogue systems and speech-to-speech translation engines, social media data mining for financial or health-related information, and the detection of attitude and emotion toward

products and services, researchers are currently developing and utilizing these technologies in practical applications [10].

4. Sentiment Analysis

1.1. Marketing, social media, and customer relationship management are just a few of the businesses that heavily rely on sentiment analysis (SA), one of the methods used in opinion mining, typically assesses a textual review's structure to see if it conveys a positive or negative impression [11]. Text organization and feature extraction, as well as parsing methods, word segmentation, POS tagging, tokenization, and word segmentation, are preprocessing processes for opinion mining. [12]. Figure 1 show Classification of sentiment analysis.

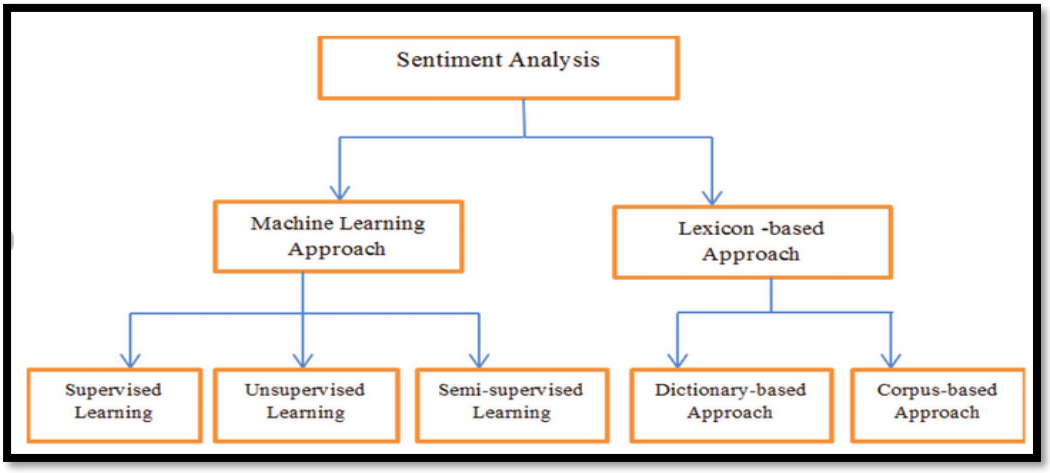

Fig.1 Classification of sentiment analysis

5. Machine Learning

A branch of Artificial Intelligence, statistics, and computer science called machine learning (ML), often known as statistical learning or predictive analytics, aims to learn from data. The use of machine learning in daily life has increased dramatically in recent years [13]. Finding patterns in data allows ML to construct computer algorithms that can adapt to new information. ML uses that data to identify patterns and change program processes [14].

6.Classification Techniques

The main premise behind classification algorithms is that they can anticipate the target class by looking at the training dataset; this prediction comes after the boundary conditions have been created. This entire process is known as classification.

6.1 Logistic Regression

Logistic Regression (LR) model is one of several supervised machine learning algorithms, based on a set of independent variables, it is used to estimate discrete values (binary values like 0/1, yes/no, and true/false), its essentially fit data to a logistic function to estimate the likelihood that an event will occur. Thus, it is often referred to as LR . As a result of predicting the likelihood as show in Equation (1), its output values fall between 0 and 1, [15].

$$y=e^{((b0+b1\ X))/(1+e^{((b0+b1\ X))}\ )} \quad \ldots(1)$$

X=input value, Y=predicted output, b0=bias or intercept term ,b1=coefficient for input (X).

6.2 Random Forest Algorithm

The Random Forest (RF) method will generate numerous training sample sets that are distinct from one another using the bagging technique. Every sample set creates a decision tree using qualities that are chosen at random [16]. A statistical machine-learning approach for prediction is called random forests. Training random choice forests, an ensemble learning strategy for classification, regression, and other applications, requires the construction of several decision trees. The class selected by the majority of trees is the output of random forests for classification tasks [17]. Each decision tree in the Random Forest is fully grown, so there is no need to cut processing. The more trees it has, the more accurate the outcome will be, and it won't over fit the data [18].

6.3 Support Vector Machine

Support Vector Machines (SVMs, also known as support vector networks) in machine learning are supervised learning models with corresponding learning algorithms that examine data used for regression and classification analysis. An SVM training algorithm creates a model from a set of training examples, each of which is marked as belonging to one of two categories, making it a non-probabilistic binary linear classifier (although there are ways to use SVM in a probabilistic classification setting, like Platt scaling). In SVM model is a mapping of the examples as points in space with as much space between the examples of each category as possible. [19]. It is also a way of supervised machine learning for classification and regression analysis [6].

6.4 Naive Bayes

Naive Bayes (NB) a probabilistic classifier built on the Bayes theory. This approach is less frequently used by researchers to make predictions. The main benefit is that, in contrast to other algorithms, it is scalable [6]. NB models have been widely used for clustering and classification. However, they are seldom used for general probabilistic learning and inference (i.e., for estimating and computing arbitrary joint, conditional and marginal distributions)[20].

7.Accuracy Measures

Loss functions, accuracy metrics, and error statistics or measures are additional means of communicating information about how well a certain forecasting technique can predict actual data [21]

7.1Prediction Measures(Metrics)

A confusion matrix is a technique for summarizing the performance of a classification algorithm, where classified data contains four sets, TP (true positive) and TN (true negative) correctly classify instances as well as FP (false positive) and FN (false negative) incorrectly classify instances .Four measures were used to assess the anticipated sentiment: precision (Prec), recall (Rec), F-1score (F1), and accuracy (Acc.) Precision and Recall are the two best-known classification metrics; they are also used for measuring the quality of information retrieval tasks in general [22]. Equations below demonstrate precision, recall, accuracy, and F1score.

$$\text{Precision} = Tp/((Tp + Fp)) \quad \ldots(2)$$

$$\text{Recall} = Tp/((Tp + Fn)) \quad \ldots(3)$$

$$\text{Accuracy} =((Tp + Tn))/(((Tp + Tn + Fp + Fn))) \quad \ldots(4)$$

$$F1 = 2*(((\text{Precision}*\text{Recall}))/((\text{Precision} + \text{Recall}))) \quad \ldots(5)$$

8. Methodology

The outline architecture of the proposed system as shown in Figure2. The first stage represents the preparation of the dataset by cleaning and feature extraction. The second stage has concentrated on the main steps of classifier model construction that after data splitting and performance evaluation.

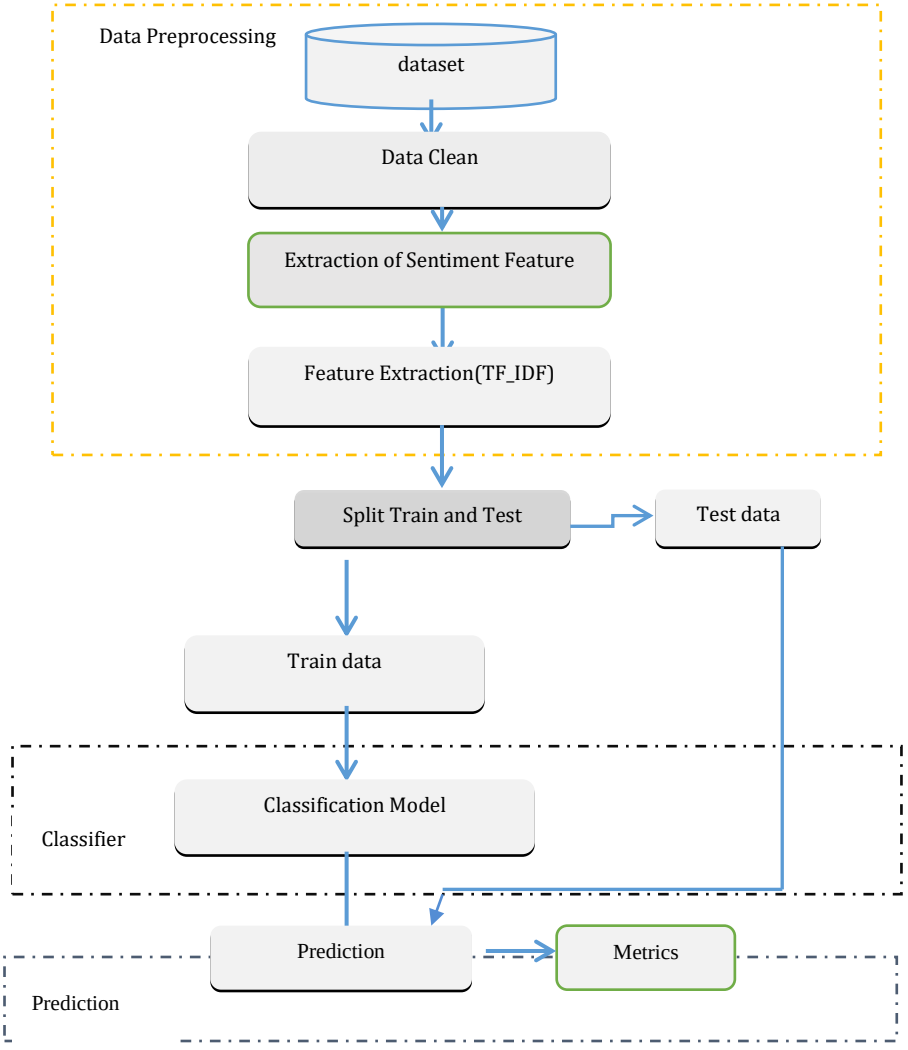

Fig. 2. Architecture of Proposed System

1. Data Set Description

The dataset utilized in this study came from the UCI ML repository's Drug Review Dataset (Drugs.com) [23], dataset have 215063 row. There are seven attributes in this dataset, the drug's name (in text), the patient's review, a 10-star patient rating (numerical) that represents how satisfied the patient is overall, the patient's health condition, the helpful count (numerical), the date (date of the review entry), and unique ID. Python is the programming language, and Anaconda Jupyter Notebook is the platform.

2.Data Cleaning

This process is known as text preparation or (preprocessing). Firstly, cleaned the reviews after removing null value, remove HTML tags, and punctuation such as ('!'$ #*? /&.;"). This is a very crucial stage in the process, URLs, etc. Eliminating words with less than two characters that do not contribute significantly to the meaning of the phrase, such as: ww, at, nm , cc, zc , hm ,er , ab . To prevent repetition, the cleaned reviews were lowercased. Stop words such "a, to, all, we, with, etc.," were also eliminated from the corpus. Additionally, the texts were tokenized, or

broken down into smaller chunks known as tokens, and then restored to their original form by performing. All the while stemming, in order to decrease text data and enhance system efficiency, a phrase's words are reduced to their stems or converted to their non-changing sections. IF a number of words have a same root, for instance (amusing, amusement, and amused), the stem would be (amus). Stemming helps reduce the amount of text data and improves the efficiency of the system. In this study, the porter stemmer was used to strip suffix ends from words like "ed," "ing," and "s," which resulted in a significant reduction in word length (dimensions). Algorithm (1) illustrated of the process of applying textual data purification to a particular paragraph.

Algorithm (1): Preprocessing on Reviews Algorithm (PoRA)

Input: List of Review unstructured

Output: List of Review structured

Begin

Step 1: For each Text Review

begin

- Convert the letter in text review to lowercase
- Remove punctuation such as ('!'$ #*? /&.;").
- Eliminating words with less than two characters such as ww, at, nm.
- Remove number from text review such as 3,4,567.
- Eliminating stop words from review such as to ,the.

end

Step2: For each word in Text Review

begin

- Tokenization text into tokens/words.
- Stemming for each word such as (amusing, amusement, and amused), the stem would be (amus).

end

Return clean review

End.

3.Sentiment Features

The Figure 3 show the number of ratings for each value from 0 to 10, which illustrates how the 10-star rating system's value counts are seen.

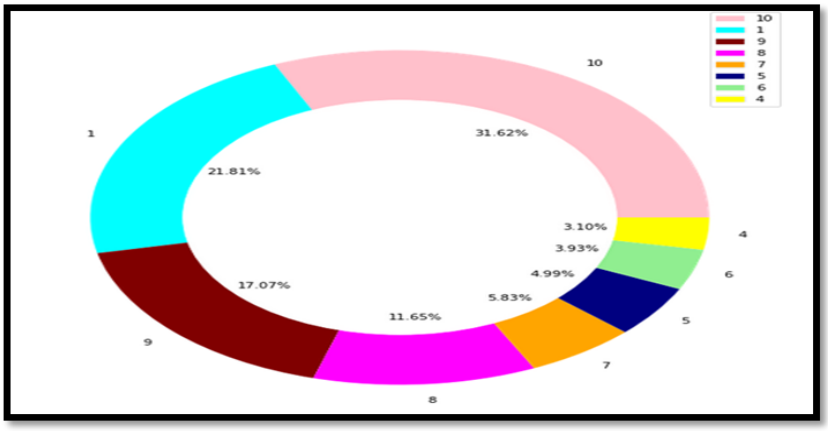

Fig.3 Statistics Analysis of Ratings

According to dataset most people choose Five rating, 10, 9, 1, 8, and 10 are more common as shown Figure 3 conclusion, the positive level is greater than the negative level. For each review in this work that was extracted as either positive or negative expressed as (1,0) called Sentiment Features (SF), it was determined by the user's star rating. Positive ratings are those with five stars or more, in contrast one to four stars are assigned to unfavorable reviews. Initially, there were 47522 negative assessments and 111583 positive evaluations in the training data, Figure 4, show frequent SF that extracted from reviews rating.

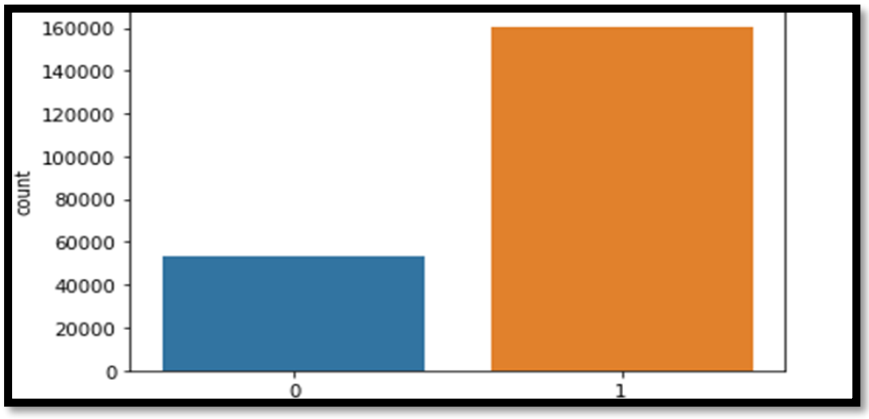

Fig.4 Count of Positive and Negative Reviews

4.Features Extraction

In this stage, the (word-document matrix) was built, the values of the cells in the matrix reflected the frequency of terms (weight of words) inside the documents, while the rows of the matrix represented the characteristics (terms). Each column in the matrix represented the number of documents (rows). (TF-IDF) in order to arrive at an estimate of the significance of the word. This matrix condenses the narratives of documents into vectors of the distinct words, which are shown as the columns of the matrix. The word-document matrix is basically used as a starting point for most algorithms.

Term frequency (TF) is a measure of how often it is that a term will be found in a document calculated as show in equation (6)

$$TF(t,d)=\log_{f0}(1+freq(t,d)) \quad \ldots(6)$$

According to Equation, the term's frequency in relation to the entire corpus is the same as the phrase's inverse document frequency (IDF). It determines the characteristics of a phrase that are unique to the text

$$IDF(t,D)=\log_{10}(N/count(d\in D:t\in d)) \quad ...(7)$$

The TF-IDF measure, which indicates how important and pertinent a word is to the content of the text, is calculated by multiplying the TF score with the IDF score, as indicated in the equation below [24].

$$T\_F\_IDF\_((t,d,D))=TF(t,d)*IDF(t,D) \quad ...(8)$$

Where N is the total number of documents ,and TF(t,d) number of features appearing in the document

5. Train Test Split

After data preparation stage according previous sub stages and smote, this dataset was divided into 20% of testing and 80% of training.

6. Classifiers

A sentiment classifier was created using various machine-learning classification methods. LR, RF, SVM, and NB.

7. Performance Metrics

Four measures were used to assess the anticipated sentiment: F1-score (F1), precision (Prec), recall (Rec), and accuracy (Acc.)

8. Results of Models

Table 2 summarizes the models classification results

Models	class	preecision	recall	F1-Score	Acc.
LR	negative	0.76	0.67	0.71	%86
	positive	0.89	0.93	0.91	
RF	negative	0.99	0.69	0.81	%92
	positive	0.91	1.00	0.95	
NB	negative	0.90	0.10	0.18	%77
	positive	0.77	1.00	0.87	
SVM	negative	0.78	0.65	0.71	%87
	positive	0.89	0.94	0.91	

9.CONCLUSION

Reviews are becoming an integral part of our daily lives, whether go for shopping, purchase something online or go to some restaurant, we first check the reviews to make the right decisions . Motivated by this, in this research sentiment analysis of drug reviews was studied different types of machine learning classifiers, such as Logistic Regression, Naive Bayes, Random Forest ,support vectors machine.

In conclusion, text mining can be useful in the medical field. In this paper, converting unstructured data to structure data by extraction sentiment features based on rating and extraction features from reviews or texts using TF-IDF technique. Supervised classification of the extraction features is trained in conjunction with sentiment features. Accuracy of Random forest classifier model is better models used as show in Table 1.

The future work of this study we can develop the work by used recommender system depended on sentiment analysis , to increase the accuracy of system used oversampling technique ,and we can used deep learning techniques rather than machine learning techniques.

References

[1] Kumar, S., De, K., & Roy, P. P. (2020). Movie recommendation system using sentiment analysis from microblogging data. IEEE Transactions on Computational Social Systems, 7(4), 915-923.

[2] Al-Ghuribi, S. M., & Noah, S. A. M. (2021). A comprehensive overview of recommender system and sentiment analysis. arXiv preprint arXiv:2109.08794

[3] Castro, F., Gelbukh, A., & González, M. (Eds.). (2013). Advances in Soft Computing and Its Applications: 12th Mexican International Conference, MICAI 2013, Mexico City, Mexico, November 24-30, 2013, Proceedings, Part II (Vol. 8266). Springer.

[4] D. Naga Swathi, Kumaran.U," An Effective Stratified K-Fold Algorithm with Logistic Regression for Drug Feedback Data", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

[5] Bemila, T., Kadam, I., Sidana, A., & Zemse, S. (2020, April). An approach to sentimental analysis of drug reviews using RNN-BiLSTM model. In Proceedings of the 3rd international conference on advances in science & technology (ICAST).

[6] Uddin, M. N., Hafiz, M. F. B., Hossain, S., & Islam, S. M. M. (2022). Drug Sentiment Analysis using Machine Learning Classifiers. International Journal of Advanced Computer Science and Applications . ,

[7] Vijayaraghavan, S., & Basu, D. (2020). Sentiment analysis in drug reviews using supervised machine learning algorithms. arXiv preprint arXiv:2003.11643.

[8] Nahma, D. R., & Abbas, A. R. (2020). Patient Opinion Mining: Analysis of Patient Drugs Satisfaction using Support Vector Machine and Logistic Regression Algorithm. Journal of Madenat Alelem College Vol, 12(2).

[9] Shiju, A., & He, Z. (2022, June). Classifying drug ratings using user reviews with transformer-based language models. In 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI) (pp. 163-169). IEEE.

[10] Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266

[11] Jain, A., Jain, V., & Kapoor, N. (2016). A literature survey on recommendation system based on sentimental analysis. Advanced Computational Intelligence, 3(1), 25-36

[12] Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. Information fusion, 36, 10-25

[13] Ricci, F., Rokach, L., & Shapira, B. (2010). Introduction to recommender systems handbook. In Recommender systems handbook (pp. 1-35) Boston, MA: springer US.

[14] Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. " O'Reilly Media, Inc.".

[15] Hilbe, J. M. (2009). Logistic regression models. CRC press

.

[16] Liu, Yingchun. "Random forest algorithm in big data environment." Computer modelling & new technologies 18.12A (2014): 147-151.

[17] Schonlau, Matthias, and Rosie Yuyan Zou. "The random forest algorithm for statistical learning." The Stata Journal 20.1 (2020): 3-29.

[18] Lin, Weiwei, et al. "An ensemble random forest algorithm for insurance big data analysis." Ieee access 5 (2017): 16568-16575

[19] Ray, S. (2018). A Comparative Analysis and Testing of Supervised Machine Learning Algorithms

[20] Lowd, D., & Domingos, P. (2005, August). Naive Bayes models for probability estimation. In Proceedings of the 22nd international conference on Machine learning (pp. 529-536)

[21]Makridakis, Spyros. "Accuracy measures: theoretical and practical concerns." International journal of forecasting 9.4 (1993): 527-529.

[22] Garg, Satvik. "Drug recommendation system based on sentiment analysis of drug reviews using machine learning." 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2021

[23] Dataset Available in / https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Druglib.com%29

[24]] Guo, A., & Yang, T. (2016, May). Research and improvement of feature words weight based on TFIDF algorithm. In 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference (pp. 415-419) IEEE.