Drug Recommendation system based on Medical Condition Classification and Sentiment Analysis of Drug Reviews

Navodya Rathnasekara, Udaya Wijenayake
Department of Computer Engineering, University of Sri Jayewardenepura, Sri Lanka

***Abstract*—The steady growth of the internet has increased the amount of user generated data on the web. In the healthcare domain, patients now commonly post their reviews about medicines after consuming them to create public awareness. Natural Language Processing techniques significantly c ontribute t o the medical field by analyzing these public reviews and identifing the effectiveness of drugs as well as understanding medical conditions they are suffering from which will help healthcare professionals and pharmacovigilance systems to ensure the physical and mental well being of the patients. Hence, this research endeavors to develop a comprehensive framework for patient medical condition classification, s entiment p rediction f rom p atients reviews and recommend suitable medicines to them. Four algorithms: Multinomial Naïve Bayes, Passive Aggressive Classifier, SGD Classifier a nd M LP C lassifier ha ve be en ap plied to medical condition classification a nd t wo a lgorithms: M ultinomial Naïve Bayes and Logistic Regression have been applied for sentiment prediction. The results demonstrate that the proposed framework has an accuracy of 94.4% for Passive Aggressive Classifier in medical condition classification a nd a ccuracy o f 9 4.85% for Logistic Regression in sentiment prediction.**

***Index Terms*—Sentiment Analysis, Classification, Recommender System, NLP, NLTK, Machine Learning**

I. Introduction

An adverse drug reaction (ADR) can be defined as a harmful or unpleasant reaction resulting from an intervention related to the use of a medicinal product. A careful medication history can assist a healthcare professional in understanding the patient's previous experiences with drug treatment, particularly in identifying previous ADRs that may preclude re-exposure to the drug [1]. Therefore, identification and monitoring of ADR is a crucial aspect of pharmacovigilance systems to ensure the safety and efficacy o f t he medicines.

Recently, social media tools have drawn more attention from researchers. Due to the increasing use of the internet, medical communities are now allowing their users (patients) to make their own comments and reviews about medicines, which are rich in useful information. These reviews of patients feedback provide enormous amounts of vital data resources to identify ADR detection. According to a survey done in Pew American Research center in 2013 [2], 59% of U.S. adults have searched online to find information about health-related topics in the past year. Also 35% of U.S. adults mentioned that they had searched online, specially to figure out what medical condition they might have. Therefore, need for medical condition prediction and a medicine recommender system is vital in the current society. Furthermore, due to the shortage of healthcare professionals, implementation of these systems can help patients to build up their knowledge on drugs and medical conditions they encounter.

Recommender system is a custom-built framework that suggests items to users based on their requirements and preferences. These frameworks use feedback from customers to analyze their sentiment and provide personalized recommendations to their specific needs. In the context of a drug recommender system, medicines will be provided to patients based on their reviews and drug ratings using sentiment analysis and feature engineering techniques within the realms of Natural Language Processing (NLP). Sentiment analysis is a series of techniques and procedures for identifying and obtaining emotional information from language including opinions and attitudes. Feature engineering is the process of creating new features from the ones that already exist which enhances the functionality of models.

The categorization procedure for sentiment analysis in the medical field was discussed by Denecke et al. [3]. The medical sentiments can be characterized by the diagnosis of medical condition, patient's health status, and medical treatment. Since it can be difficult to collect context-relevant terms that indicate the polarity of sentiment, sentiment analysis in the medical field is a challenging task. Furthermore, the language used by the patients may vary greatly and have different meanings [4].

In this paper, the role of medical condition classification and sentiment analysis in the healthcare industry is discussed. Since patients' clinical reports tend to contain less subjective information, reviews written by patients after taking the medication were analyzed. Patients anonymously provide their reviews of different medicines, describing how the drug affects their ability to fight the illness. These evaluations will help the healthcare professionals and pharmacovigilance systems to enhance the mental and physical well-being of the patients. The major contribution of the work is summarized below:

- Classification of the patients' medical condition based on the review they give.

Correspondence: Navodya Rathnasekara (E-mail: prabuddhi.rath2@gmail.com)
Received: 16-06-2024 **Revised:** 12-08-2024 **Accepted:** 09-09-2024
Navodya Rathnasekara and Udaya Wijenayake are from University of Sri Jayewardenepura (prabuddhi.rath2@gmail.com, udayaw@sjp.ac.lk)
DOI: https://doi.org/10.4038/icter.v18i2.7291
The 2025 Special Issue contains the full papers of the abstracts published at the 24th ICTer International Conference.

- Sentiment prediction (positive, negative) of the patients' review.
- Drug recommendation system according to the medical condition and sentiment score.

The study is divided into seven parts: The Introduction section offers a brief overview of the research's necessity; the Literature Review section offers a summary of earlier studies conducted in this domain; the Methodology section details the techniques used in this study; The Result section evaluates applied model results using various metrics; the Discussion section outlines the significance of the findings; Future work section outlines framework's limitations, and lastly, the conclusion section summarizes the study.

II. Literature Review

Multilingual sentiment analysis was carried out in this study [5] by applying Recurrent Neural Networks (RNN) and Naive Bayes algorithms. Multilingual tweets were translated into English using the Google Translate API, and results show that RNN performed 95.34% of accuracy compared to Naive Bayes, which was 77.21%.

Drug reviews gathered from the Drugs.com website, which offers information on drugs, were assembled into a dataset by Grasser et al. [6] . Every drug review has a score, ranging from 0 to 9, that represents the patients' level of satisfaction with the medication. Based on their ratings, the reviews were divided into three classes: negative (rating ≤ 4), neutral (rating in $[5, 6]$), and positive (rating ≥ 7). With a logistic model, authors achieved an accuracy of 0.9224.

Satvik Garg [7] proposed a drug recommender system using several algorithms and vectorization techniques where Linear SVC on TF-IDF vectorization outperformed all other models with 93% of accuracy. In contrast, the Decision tree classifier on Word2Vec showed the worst performance by achieving only 78% of accuracy. Emotion values for each algorithm were as follows: Perceptron on Bow (91%), Linear SVC on TF-IDF (93%), LGBM on Word2Vec (91%), and Random Forest on manual features (88%).

Zhang Min [8] proposes a Weakly Supervised Mechanism (WSM) that applies the weakly labeled data to pre-train the parameters of the model and then uses labeled data to fine-tune initialed parameters. The approach reduces the effect of noise data on the consequences. After that, WSM combines a Convolutional Neural Network (CNN) and a Bi-directional Long Short-term Memory (Bi-LSTM) named as WSM-CNN-LSTM to complete the task of sentiment classification on ADR.

Most of the above-mentioned research focuses on the sentiment classification of the review to find if the given review is positive or negative. However, in this research, we extend our focus beyond sentiment analysis alone to implement a more comprehensive approach that integrates both sentiment analysis and medical condition classification from the patients' reviews. So the proposed system has mainly two parts.

- Classify the patients' medical condition based on the review they have given.
- Predict the sentiment (positive, negative) of the patients' review.

This helps the users who do not have a proper understanding about their medical situation to identify their condition by bridging the existing gap in literature and to identify the effectiveness of their medicines from the sentiment classification.

III. Methodology

A. Dataset

The Drug Reviews Dataset is taken from the UCI Machine Learning Repository. It contains patient reviews on drugs and the related medical conditions with a 10-star patient rating. This data was obtained by web scraping online pharmaceutical review sites. The Drug Review Data Set contains approximately 215063 data with 7 columns including the patients' 'review', 'drug Name' (the name of the drug), 'condition' (the medical condition the patient is suffering from), 'rating' (10-star patient rating for the drug), 'date' (the date of the entry), 'uniqueId' and the 'usefulCount' (number of users who found the review useful). These reviews are from the year 2008 to 2017. Here, the sentiment of the review is the target variable for the sentiment prediction, and condition is the target variable for medical condition classification. In the dataset, sentiment of any review is not given, therefore it is necessary to assign sentiment based on the rating first and then use it as the target variable. In the dataset, the drugName and condition are categorical features, date is a date object, rating and usefulCount are numerical features, and review is a text feature.

Figure 1 shows the proposed model used to build the drug recommender system. It contains four stages, namely, data preparation, classification, evaluation, and recommendation.

B. Data Cleaning and Pre-Processing

This research applied standard data cleaning and preprocessing techniques like checking null values, and duplicate rows, and removing unnecessary values, and text from rows. Therefore, all 1194 rows with null values in the conditions column were removed, as shown in Figure 2. It was also considered that the 'uniqueId' should be unique in order to remove duplicate values.

Then, the reviews are cleaned before the feature engineering. Regular expressions are used to clean the reviews. The reviews are first converted to lowercase, and unnecessary special characters and html characters are removed from the text. Secondly, non-alphabetical characters are removed from the reviews. Finally, trailing and leading whitespaces are removed from the reviews. Multiple whitespaces are replaced with a single space for better clarity.

The cleaned reviews are then tokenized to convert the texts into small pieces called tokens. Also, stop words, for example, "a", "the", "with", "is", "are" etc.," are removed from the corpus using *nltk.corpus.stopwords*. Then, the tokens are returned to their base forms by performing lemmatization on all tokens using *nltk.stem.WordNetLemmatizer*.

C. Exploratory Data Analysis

Exploratory data analysis is mostly used to improve understanding of the variables in the dataset and their relationships,

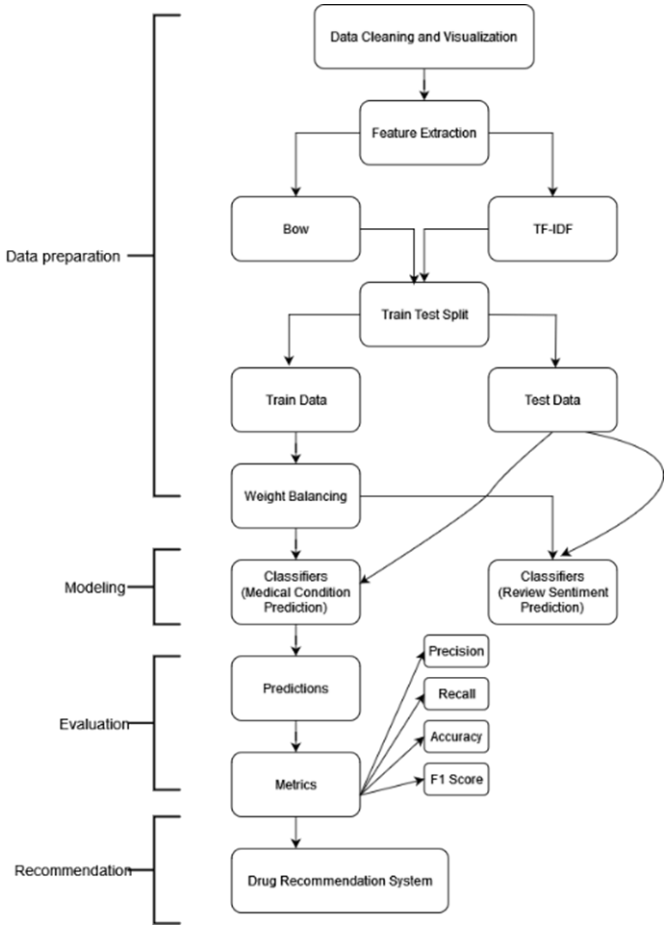

Fig. 1: Flowchart of the proposed model

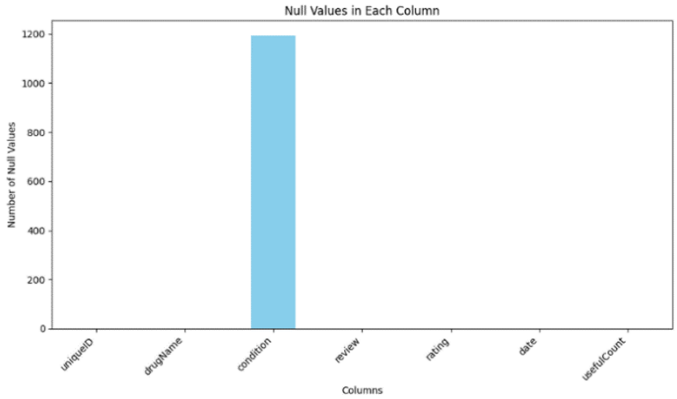

Fig. 2: Null Values in Each Column

as well as to discover what the data could reveal beyond the formal modeling or hypothesis testing work. It can also assist with deciding the suitability of the statistical methods which used for data analysis. Since there are many features (uniqueID, drugName, condition, review, rating, date, usefulCount) in the dataset, it was crucial to understand the key features and how they are related to one another.

According to Figure 3, it can be shown that the dataset is highly imbalanced and the conditions such as Birth Control, Depression, Pain have more reviews than something like ADHD or High Blood Pressure.

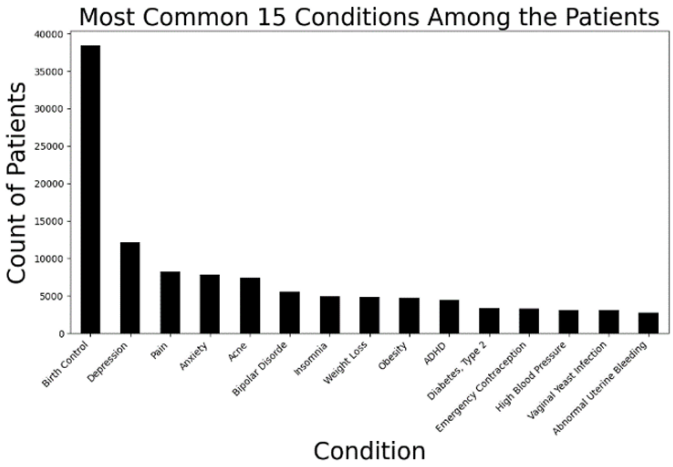

Fig. 3: Data Count for the Most Common 15 condition

A word cloud is a visually engaging way to represent text data, which depicts the keyword metadata (tags) on websites, or provides visual representation of free form text. Figure 4 illustrates the word cloud for condition 'Acne' where words such as skin, acne, and face are more frequently used in the reviews.

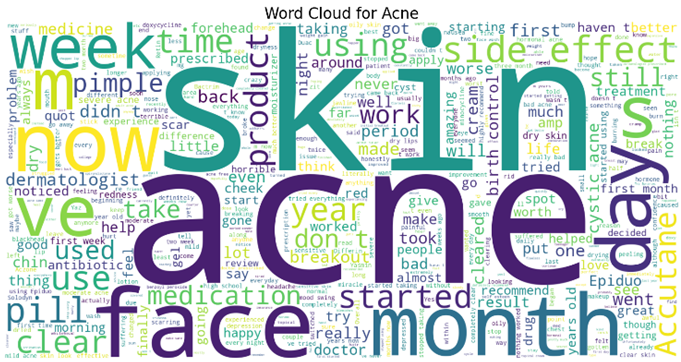

Fig. 4: Word Cloud for Acne

Figure 5 illustrates the distribution of the ratings from 1 to 10 in the dataset. It shows that 10/10 has the highest rating count and after that 9 and 1. Data points of rating of the drugs from 2 to 7 is low.

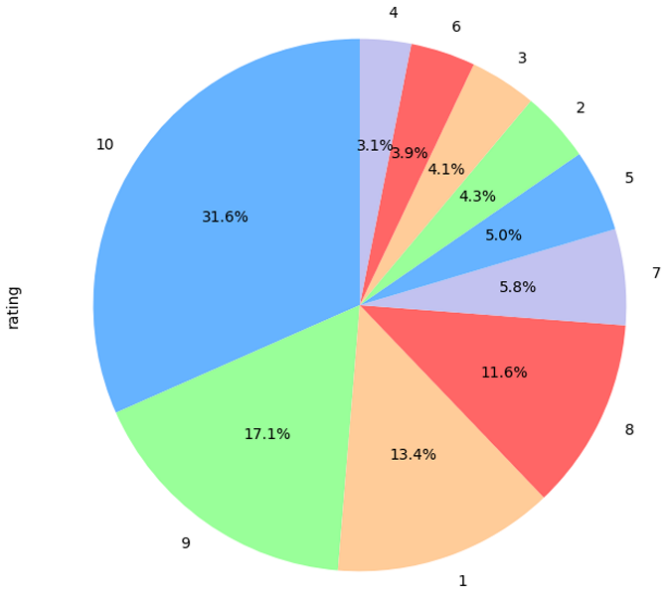

Fig. 5: Distribution of ratings

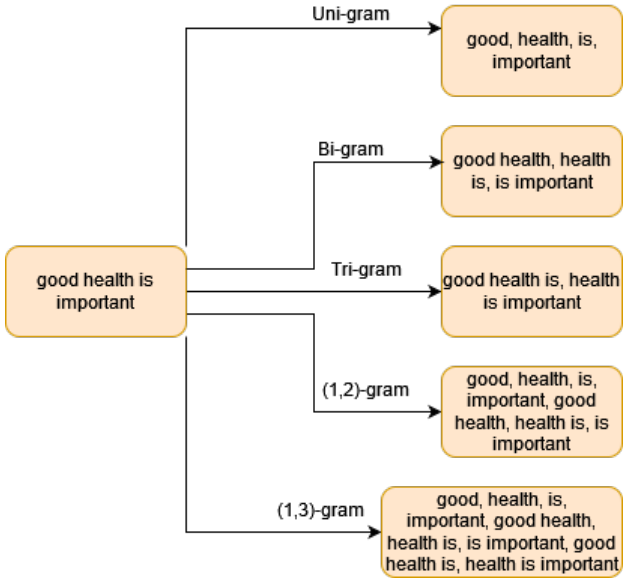

Fig. 6: Comparison of various grams in a sentence

D. Feature Extraction and Model Development

1) Vectorization

a) Bow: Bag of words is an algorithm used in natural language processing that counts the frequency each token appears in a review. A token can be referred to as a single word (a unigram) or as any arbitrary number of words (n-grams). Figure 6 outlines how unigrams, bigrams, and trigrams framed from a sentence [7].

b) TF-IDF: In this technique, words are assigned with weights rather than counts. The idea is to give low priority to the words that occur frequently in the dataset, indicating that TF-IDF measures relevance rather than recurrence. Also, TF (Term Frequency) represents the probability of finding a word in a document.

$$\text{tf}(t,d) = \log(1 + \text{freq}(t,d)) \tag{1}$$

Inverse Document Frequency (IDF) is the opposite of the number of times a specific word showed up in the whole corpus.

$$\text{idf}(t,d) = \log\left(\frac{N}{\text{count}(d \in D : t \in d)}\right) \tag{2}$$

TF-IDF is the multiplication of TF with IDF, suggesting how important and relevant a word is in a document.

$$\text{tfidf}(t,d,D) = \text{tf}(t,d) \cdot \text{idf}(t,D) \tag{3}$$

2) Data Reduction

Before applying suitable classifiers, it is important to perform data reduction by removing unnecessary columns in the dataset as not every column is needed for model training. It helps to improve the efficiency, interpret-ability, and generalization performance of classifiers. Therefore, for the medical condition classification only the condition and review column are used. For sentiment prediction rating and review columns are used, where rating¿5 was considered as positive sentiment and others as negative sentiments.

There were 916 unique conditions in the dataset, and the dataset was highly imbalanced. For medical condition prediction, there should be sufficient data points for each class; otherwise the model accuracy would be very low. Therefore, only the top 15 conditions were selected for the medical condition classification. Hence, there were 114308 data points in the dataset for condition classification. But sentiment prediction did not have that problem.

3) Train Test Split

The dataset was split into 80% for training and 20% of testing. When splitting the data, an equal random state was set to ensure reproducibility of the generated datasets. Then a weight balancing technique was applied because the dataset was highly imbalanced.

4) Classifiers

Distinct machine-learning classification algorithms were used to build the medical condition classifier and sentiment prediction. Multinomial Naive Bayes, Passive Aggressive Classifier, Stochastic Gradient Descent Classifier and Multilayer Perceptron Classifier were used with Bow and TF-IDF vectorizers to predict the medical condition. Multinomial Naive Bayes and Logistic regression were used with Bow and TF-IDF vectorizers to predict the sentiment of the review. Also, grid search was used to find the best combination of hyper-parameter values for each model.

5) Metrics

The model results were measured using four metrics, namely, Precision, Recall, F1score and Accuracy. If Tp = True positive (occurrences where model predicted the positive sentiment truly), Tn = True negative (occurrences where model predicted the negative class truly), Fp = False positive (occurrences where model predicted the positive class falsely), Fn = False negative (occurrences where model predicted the negative class falsely):

$$\text{Precision} = \frac{T_p}{T_p + F_p} \tag{4}$$

$$\text{Recall} = \frac{T_p}{T_p + F_n} \tag{5}$$

$$\text{Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{6}$$

$$F1score = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

IV. Results

Table I shows the results using Accuracy on Bow and TF-IDF vectorization techniques for medical condition classification. Among them, Passive Aggressive Classifier has the highest accuracy of 94.4% for TF-IDF. Both SGD Classifier and Passive Aggressive Classifier showed similar type of results for both Bow and TF-IDF. Multinomial Naive Bayes showed lower accuracy with TF-IDF technique compared to Bow.

Also, when considering Passive Aggressive Classifier and SGD Classifier in TF-IDF, the accuracy increase from Unigram to Bigram is considerable. However, from Bigram to Trigram, the increase of accuracy is negligible. Thus, Multinomial Naïve Bayes has showed considerably lower accuracy in TF-IDF.

TABLE I: Results for Medical Condition Prediction.

Classifier	Accuracy(%) - Bow			Accuracy(%) - TF-IDF		
	Uni gram	Bi gram	Tri gram	Uni gram	Bi gram	Tri gram
Multinomial Naïve Bayes	84.9	83.5	83.9	70.7	49.9	46.4
Passive Aggressive Classifier	87.4	93.1	92.8	89.7	94.2	94.4
SGD Classifier	85.9	93.6	93.9	88.4	93.6	93.5

	precision	recall	f1-score	support
ADHD	0.96	0.94	0.95	902
Abnormal Uterine Bleeding	0.96	0.86	0.90	549
Acne	0.98	0.97	0.97	1487
Anxiety	0.91	0.86	0.88	1563
Bipolar Disorde	0.94	0.91	0.92	1121
Birth Control	0.98	0.99	0.98	7687
Depression	0.90	0.93	0.91	2433
Diabetes, Type 2	0.97	0.96	0.97	672
Emergency Contraception	0.99	0.99	0.99	658
High Blood Pressure	0.93	0.90	0.91	621
Insomnia	0.88	0.92	0.90	981
Obesity	0.87	0.84	0.86	951
Pain	0.96	0.95	0.95	1649
Vaginal Yeast Infection	0.97	0.97	0.97	617
Weight Loss	0.86	0.87	0.87	971
accuracy			0.94	22862
macro avg	0.94	0.92	0.93	22862
weighted avg	0.94	0.94	0.94	22862

Fig. 7: Classification report of Passive Aggressive Classifier for medical condition prediction

TABLE II: Accuracy Of MLP for Medical Condition Prediction.

Number of Hidden Layers	Accuracy (%) - Bow +Trigram
3	82.4
5	91.5
7	92.7
9	93.2

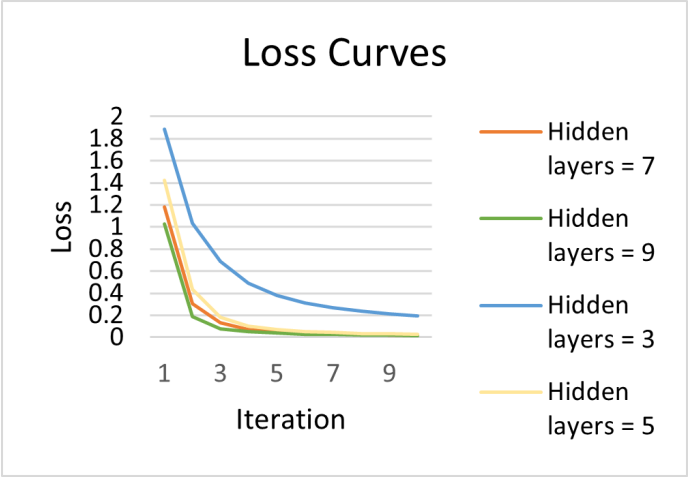

Fig. 8: Comparison of loss curves in MLP

Other than the machine learning classifiers, a deep learning approach using Multilayer Perceptron Classifier was used to train the medical condition prediction model. It is a feed-forward neural network with input layer, output layer and one or more hidden layers. Table II shows the accuracy of MLP with number of hidden layers. According to that, the accuracy increases when the number of hidden layers is increasing, but the rate of improvement in accuracy diminishes as the network becomes deeper.

After the medical condition classification, the next part is sentiment prediction of the review. According to Table III, Logistic Regression has the highest accuracy for sentiment prediction. Also, both algorithms performed better in both Bow and TF-IDF vectorization techniques.

TABLE III: Accuracy for Sentiment prediction.

Classifier	Accuracy (%) Bow + Trigram	Accuracy (%) TF-IDF Trigram
Multinomial Naïve Bayes	92.47	92.61
Logistic Regression	94.85	88.58

In order to recommend suitable medicines for a predicted medical condition, the top 3 drugs are extracted based on ratings and useful count of that particular drug. Figure 9 and Figure 10 show the output of the web application for two different reviews. It can be seen that drugs are recommended only if the sentiment is negative.

V. Discussion

According to the results, Naive Bayes model has performed better when the number of classes in the dataset is less. When

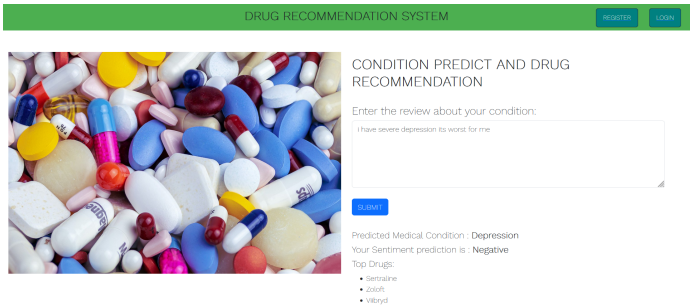

Fig. 9: Output for negative sentiment

Fig. 10: Output for positive sentiment

the medical condition prediction has 15 classes, Naive Bayes shows less accuracy. But sentiment prediction with Naive Bayes shows high accuracy because it has only 2 classes (positive/negative). The reason for this result is the class imbalance within the dataset. When there are 15 classes in the condition prediction, some classes have a very smaller number of data and makes it challenging for the classifier to perform well on those classes.

Also, according to the results, simple machine learning classifiers like Passive Aggressive Classifier and SGD Classifier performed better than deep neural networks like MLP. The reason for this is that there is less amount of data in the dataset. Deep learning models require large amount of data because of the complexity and capacity of those models. Since there is many hidden layers in the deep neural networks, it requires a considerably large amount of data to train the parameters. Also when there is less number of data these models tend to over-fit as, it learns noisy data rather than the generalized patterns of the dataset.

Models like Passive Aggressive Classifier and SGD Classifier performs better on small datasets as they have fewer parameters to tune and are less prone to over-fitting. These classifiers update their models incrementally with each new training sample and making it faster and efficient.

VI. Future Work

Future work involves training the models to predict more medical conditions than the current system so it can be used efficiently in real world applications. By incorporating more medical conditions, the system can support more patients and healthcare professionals.

Also building sentiments and lexicons which are specific to medical domain is an essential area of future research. The current sentiment analysis tools lack of medical related specificity and nuances. Therefore, this approach will contribute significantly to understanding the sentiments and classification tasks within the healthcare context. In addition, a multilingual drug recommendation system can be implemented using data driven approaches.

VII. Conclusion

Reviews are becoming a vital part in our daily lives, especially in the e-commerce sector. Analyzing user reviews is important to determine user satisfaction and improve products or services. In the healthcare domain, analyzing patients' medical conditions and the Adverse Drug Reactions they experience is crucial as these factors affect their physical and mental well-being.

Motivated by this, this research aims to implement a comprehensive framework to build a drug recommender system by analyzing drug reviews from patients using classification techniques and sentiment analysis. Five algorithms: (Multinomial Naive Bayes, Passive Aggressive Classifier, SGD Classifier, Logistic Regression, Multilayer Perceptron Classifier) were used with Bow and TF-IDF vectorization techniques to train the models. The system showed 94.4% accuracy for medical condition classification using TF-IDF vectorizer and Passive Aggressive Classifier. For sentiment prediction, accuracy was 94.85% for Bow with Logistic regression model. However, unequal class distribution and a small dataset were a major issues in training and testing of the models which led to lower accuracy.

References

[1] J. J. Coleman and S. K. Pontefract, "Adverse drug reactions," *Clinical Medicine*, vol. 16, no. 5, pp. 481–485, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470211824024928

[2] S. Fox, "Health and technology in the u.s." Dec 2013. [Online]. Available: https://www.pewresearch.org/internet/2013/12/04/health-and-technology-in-the-u-s/

[3] K. Denecke and Y. Deng, "Sentiment analysis in medical settings: New opportunities and challenges," *Artificial Intelligence in Medicine*, vol. 64, no. 1, pp. 17–27, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365715000299

[4] S. Gohil, S. Vuik, and A. Darzi, "Sentiment analysis of health care tweets: Review of the methods used," *JMIR Public Health Surveill*, vol. 4, no. 2, p. e43, Apr 2018. [Online]. Available: http://publichealth.jmir.org/2018/2/e43/

[5] V. Goel, A. K. Gupta, and N. Kumar, "Sentiment analysis of multilingual twitter data using natural language processing," in *2018 8th International Conference on Communication Systems and Network Technologies (CSNT)*, 2018, pp. 208–212.

[6] F. Gräßer, S. Kallumadi, H. Malberg, and S. Zaunseder, "Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning," in *Proceedings of the 2018 International Conference on Digital Health*, ser. DH '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 121–125. [Online]. Available: https://doi.org/10.1145/3194658.3194677

[7] S. Garg, "Drug recommendation system based on sentiment analysis of drug reviews using machine learning," in *2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, 2021, pp. 175–181.

[8] Z. Min, "Drugs reviews sentiment analysis using weakly supervised model," in *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2019, pp. 332–336.