IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DRUG RECOMMENDATION SYSTEM USING NLP AND MACHINE LEARNING APPROACH

[1]Punidha A, [2]Arjun Balaji, [3]Jaisri V, [4]Mithra Saravanan

[1]Assistant Professor (Sr. G.)

Department of Artificial Intelligence and Data Science

KPR institute of engineering and Technology, Coimbatore, Tamil Nadu.

Abstract: A drug recommendation system is a technological solution that helps doctors prescribe the right drugs to patients based on their demographics, history of illness, symptoms, and drug efficacy, among other things. The system analyses massive datasets, including patient data, medication profiles, clinical research, and drug-drug interactions, among others, by using sophisticated algorithms and methodologies. This innovative approach analyses user sentiment and emotions in text data by using big data and sophisticated analytics. As a result, a drug recommendation system is now required to assist doctors and patients in understanding how different medications may affect their medical conditions. A software programmed that makes product recommendations based on user wants and preferences is called a recommender framework. It makes utilization of client surveys to anticipate customized remedies and comprehend feelings. Sentiment analysis and feature engineering are used by drug recommender systems to identify patients with certain conditions and provide the right medication.

Keywords: Drug Recommendation, medicine, health, symptoms, healthcare.

I. INTRODUCTION:

Due to a surge in diseases, particularly in rural areas with fewer specialists, many countries are facing a shortage of physicians. The lengthy six to twelve-year training period for doctors makes it challenging to quickly address this issue. In response to this crisis, there is a need to maximize the utilization of Telemedicine frameworks. Currently, clinical errors are prevalent, with prescription errors causing harm to over 300 thousand people in China and 150,000 people in the USA annually. Medical experts make mistakes in more than 42% of prescriptions due to their limited expertise. It is crucial to prioritize the selection of the best prescription, necessitating experts with a wealth of patient knowledge, antibacterial drugs, and microscopic organisms.

As the internet and online commerce continue to advance rapidly, the significance of product reviews has become undeniable in global purchasing decisions. Consumers worldwide have grown accustomed to scrutinizing reviews and websites before making purchasing decisions. While there has been extensive research on predicting ratings and recommendations in the e-commerce sector, the healthcare or medical treatments domain has been relatively overlooked. Despite this, there is a noticeable increase in individuals seeking health information online and attempting to self-diagnose through digital platforms. This trend underscores the significance of delving deeper into this area. A 2013 research study revealed that approximately 65% of adults used the internet to search for health-related information, with around 40% attempting to diagnose health conditions online. Consequently, the development of a medication recommender system becomes essential to support healthcare professionals and empower patients in understanding medications relevant to specific health conditions. The incorporation of sentiment analysis and feature engineering involves employing various techniques, methodologies, and tools to identify and extract emotional information, including opinions and attitudes, from language. Sentiment analysis focuses on deciphering emotional data, while feature engineering boosts model performance by creating additional features derived from existing ones.

II. BACKGROUND:

2.1 Motivation:

The objective of this study is to elucidate the optimal sequence of steps leading to the development of an effective drug recommendation system. Drawing from various sources, extracting pertinent data from medical records, analyzing the retrieved information, and generating personalized medication suggestions constitute the primary functions of such a system. The quality of each task is contingent upon the proficiency of the preceding activities. Within the realm of Natural Language Processing, drug recommendation systems employ text summarization techniques to distill complex medical information into concise summaries. By amalgamating Machine Learning and Data Mining methodologies, drug recommendation systems aim to streamline the decision-making process for healthcare professionals and patients. At its core, the overarching objective of a drug recommendation system is to identify a subset of data that encapsulates all relevant information from the broader dataset, thereby facilitating informed medication choices.

2.2 Action to implement:

The actions to be implemented for drug recommendation system,

- Data collection and preprocessing are paramount. This involves gathering relevant medical data from various sources such as patient records, clinical trials, and scholarly articles. The collected data must then undergo preprocessing to ensure uniformity and cleanliness, including tasks such as text normalization, tokenization, and data cleansing.
- Utilizing techniques such as sentiment analysis, feature engineering, and collaborative filtering, the system can extract meaningful insights from the preprocessed data. By training machine learning models on historical patient data and medication outcomes, the system can learn patterns and correlations to make accurate medication recommendations.
- The deployment of the developed system into a user-friendly interface accessible to healthcare professionals and patients which entails the medication recommendation in a comprehensible manner which additionally suggests its effectiveness and adaptability to changing healthcare needs.

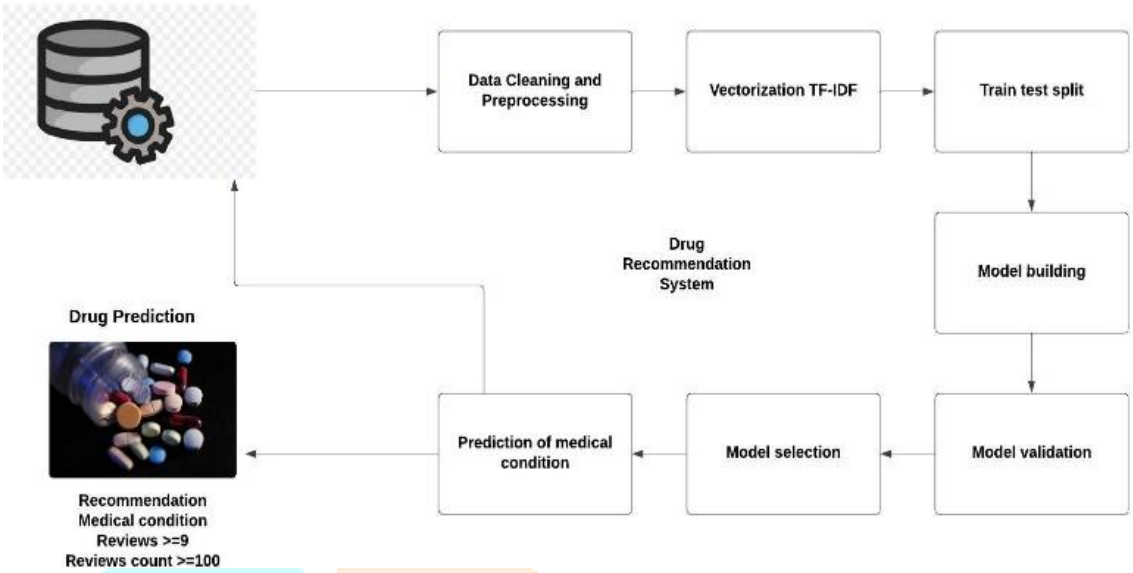

Fig 1. Overall Flowchart

2.3 Need of Drug Recommendation System:

In the realm of modern healthcare, the necessity for a drug recommendation system is becoming increasingly evident. This system serves as a critical tool in assisting healthcare providers and patients in making informed decisions regarding medication prescriptions and treatments. With the vast amount of medical information available, ranging from drug reviews to patient feedback, there exists a pressing need to streamline this data into actionable insights. A vast dataset can be combed through by a drug recommendation system using machine learning and sentiment analysis to find trends, preferences, and possible adverse effects linked with different prescriptions. This not only aids healthcare professionals in optimizing treatment plans but also empowers patients to actively participate in their healthcare journey. Furthermore, by facilitating personalized and evidence-based recommendations, such a system has the potential to improve patient outcomes, enhance medication adherence, and ultimately contribute to the overall efficiency and effectiveness of healthcare delivery.

2.4 Business Objective of Drug Recommendation System:

The primary business objective of a drug recommendation system is to enhance the quality of healthcare services by providing personalized medication recommendations tailored to individual patient needs. By leveraging advanced algorithms and data analytics, the system aims to optimize treatment outcomes while minimizing adverse effects and risks associated with medication use. Additionally, the system seeks to improve patient satisfaction and engagement by empowering individuals to make informed decisions about their health. From a business perspective, the implementation of a drug recommendation system can lead to increased efficiency in healthcare delivery, reduced healthcare costs, and enhanced patient retention. Moreover, by fostering a more collaborative and patient-centered approach to healthcare, the system can contribute to the overall competitiveness and reputation of healthcare providers in the market. Ultimately, the business objective of the drug recommendation system is to drive value by improving patient outcomes, enhancing the patient experience, and fostering long-term relationships with healthcare consumers.

III. METHODOLOGIES:

The Drug Recommendation System employs machine learning techniques for sentiment analysis, enabling the creation of an accurate medication recommendation system. A comprehensive dataset, comprising drug reviews and textual information, is collected from reputable sources such as medical forums, social media platforms, and healthcare databases. Prior to analysis, the dataset undergoes preprocessing to remove irrelevant information and ensure uniformity. The dataset utilized in this research is sourced from the UCI Machine Learning repository, named the Drug Review Dataset. It includes six attributes: drug name, patient reviews, patient condition, beneficial count, which is a numerical indicator of how many people found the review useful, review entry date, and 10-star patients rating, which is a numerical indicator of the general patient satisfaction. In total, the dataset comprises 215,064 instances.

The medication recommendation system is currently under development, employing collaborative filtering, content-based filtering, and hybrid filtering methodologies simultaneously. Collaborative filtering examines user preferences and behaviors, while content-based filtering matches medication attributes with user profiles. Additionally, sentiment analysis findings are incorporated into the model to enhance recommendations. The algorithm prioritizes medications with favorable sentiment ratings, considering potential side effects and concerns highlighted in unfavorable reviews. To evaluate performance, the Drug Recommender System utilizes a test dataset containing user preferences and feedback, assessing metrics such as precision, recall, accuracy, and F1-score. Users' input is analyzed to gauge utilization and satisfaction levels. The overarching aim is to deliver reliable, personalized, and sentiment-aware drug recommendations, empowering healthcare providers and patients to make informed decisions regarding prescriptions and treatments, ultimately enhancing patient outcomes and healthcare experiences.

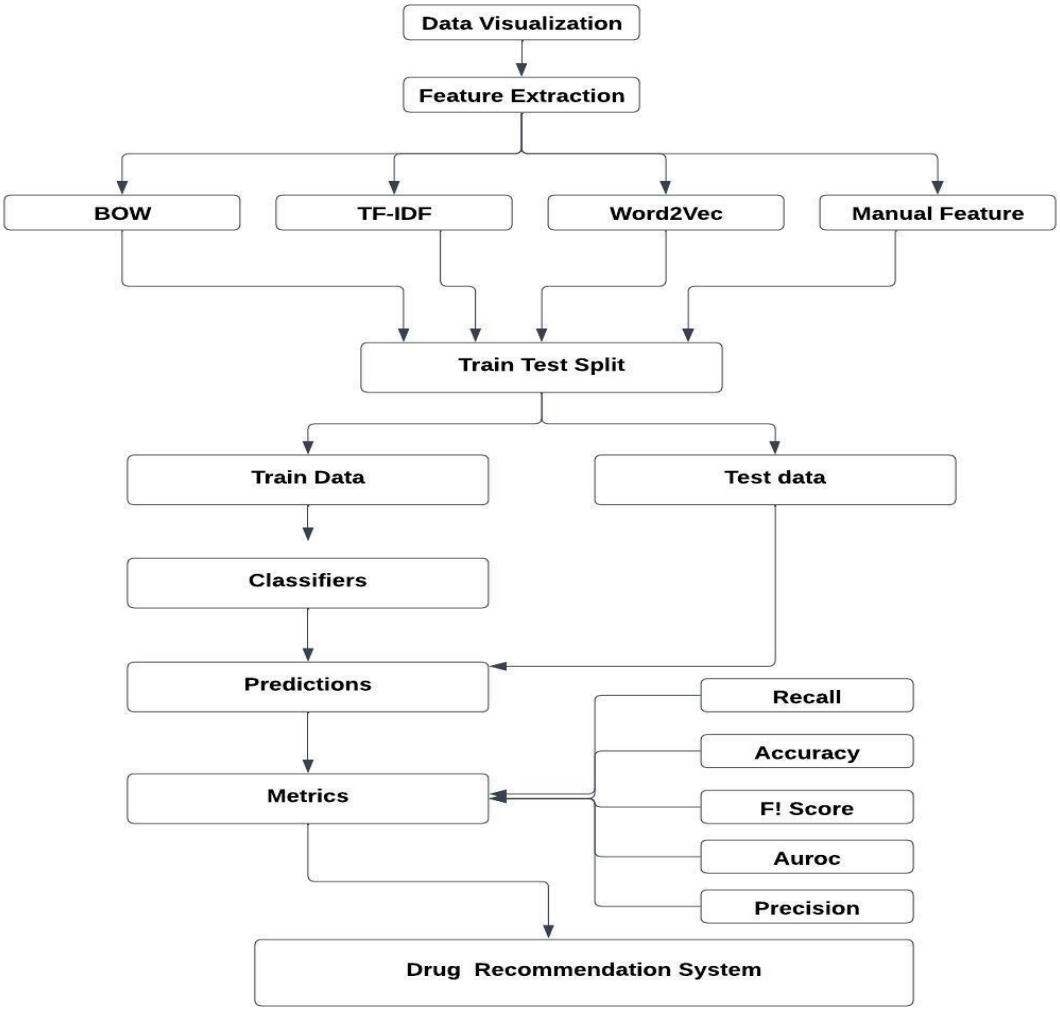

Fig 2. Flowchart of the proposed model

3.1 Data Cleaning:

Data preparation techniques were employed, including the identification and handling of null values, duplicate rows, and unnecessary data. Additionally, text was extracted from rows as part of the preprocessing process. Specifically, 1200 null value rows were identified and subsequently removed from the conditions column to ensure data integrity and accuracy.

The features, namely Given that the terms for both diseases and drugs are thought to have predictive power, the health problem and medication columns were included into the review text. However, before moving on to the feature extraction phase, it is crucial to preprocess the review text for accurate vectorization. This preprocessing involves cleaning the reviews by removing HTML tags, punctuation marks, quotes, URLs, and other extraneous elements. Subsequently, the cleaned reviews were converted to lowercase to avoid redundancy, followed by tokenization to break the text into smaller units known as tokens. Additionally, stop words such as "a," "to," "all," "we," and "with" were eliminated from the corpus to improve data quality. The tokens were then lemmatized to bring them back to their base forms, standardizing the text for analysis. Furthermore, for sentiment analysis, each review was labeled as either positive or negative based on the user rating. Reviews with ratings ranging from 6 to 10 were categorized as positive, while those with ratings below 6 were categorized as negative.

3.2 Feature Extraction:

Following text preprocessing, the data needed to construct sentiment analysis classifiers required careful setup. Since machine learning algorithms cannot directly process text, it must be transformed into numerical format, typically represented as vectors. In this investigation, widely recognized techniques for extracting features from text data were applied, encompassing the bag of words, TF-IDF, and Word2Vec methods. Furthermore, manual feature engineering methods were employed to manually extract features from the review column, resulting in the development of an alternative model termed the manual feature model. This model stands apart from the BoW, TF-IDF, and Word2Vec approaches.

Bow: The Bag of Words algorithm, employed in natural language processing, operates by counting the occurrences of tokens within a given review or document. Tokens can be single words or a variable number of words, known as n-grams. In this study, the n-gram range selected was (1,2), allowing consideration of both single words and pairs of consecutive words. However, a notable limitation of the BoW model is its equal treatment of all terms, regardless of their frequency within the corpus. This approach can lead to the generation of a large matrix during training, resulting in computational inefficiencies. The model may become computationally burdensome, especially when dealing with extensive datasets, potentially impacting performance and resource utilization. Consequently, it is essential to address this limitation and explore alternative approaches that account for the varying importance of terms within the corpus during the feature extraction process.

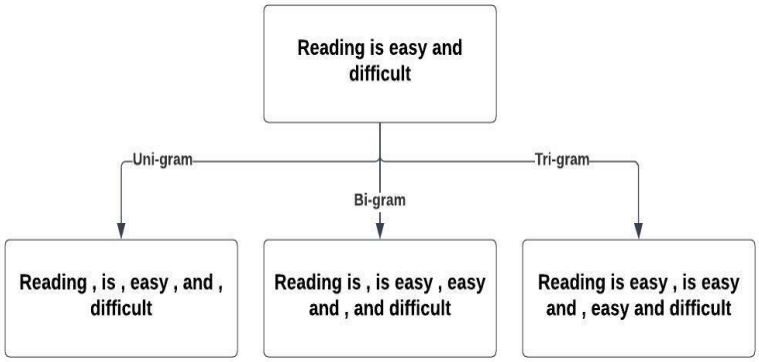

Fig 3. Types of grams framed from a sentence.

- **TF-IDF:** TF-IDF (Term Frequency-Inverse Document Frequency) serves as a statistical metric within natural language processing, assessing the significance of a word within a document in comparison to a broader corpus. The calculation involves the multiplication of the term frequency, representing how frequently a term occurs in a document, by the inverse document frequency, which penalizes terms occurring frequently across the entire corpus. The TF-IDF formula is expressed as follows:

  TF-IDF = TF (term, document) * IDF (term, corpus)

  Where TF (term, document) = (Number of times term appears in document) / (Total number of terms in document) And IDF (term, corpus) = log (Total amount of documents / Number of documents containing term)

- **Word2Vec:** Word2Vec is a model specifically crafted for generating word embeddings through the utilization of diverse deep learning techniques. Operating on extensive corpora, Word2Vec transforms input from large text datasets into a vector space, typically characterized by numerous dimensions. The fundamental idea is to grasp the semantic essence of words and arrange word vectors in a manner that aligns words with akin meanings in proximity within the vector space.

- **Manual features:** Feature engineering, a widely embraced practice, plays a pivotal role in elevating model accuracy. In this study, we integrated fifteen features into our model. These encompassed the useful count, the condition column which underwent label encoding via the label encoder function from the Scikit library, along with day, month, and year attributes derived from the date column using the Date Time function from the panda's library. Moreover, the Text Blob toolkit was employed to extract polarity values from both cleaned and uncleaned reviews, subsequently augmenting them as features. Additionally, a total of eight features were generated from each of the text reviews to enhance the overall feature set of the model.

3.3 Train Test Split:

Four Sets of data were generated utilizing Bow, TF-IDF, Word2Vec, and manual features. Subsequently, each dataset underwent a division, allocating 75% for training and 25% for testing purposes. To ensure uniformity, the data split was executed with an identical random state, guaranteeing consistent generation of random numbers for the train-test split across all four datasets.

Table 1: Enumerate the manually extracted features from user reviews.

Features	Description
Title	Calculates the word count in the title
Letter	Calculates the letter count
Word	Determine the word count.
Unique	Calculates the count of distinct words.
Stopwords	Determine the frequency of stopwords.
Punctuation	Calculate the occurrence of punctuation
Upper	Number of words written in uppercase.
Average	Calculate the average length of words.

3.4 Algorithms Implemented:

Passive Aggressive Classifier: The Passive Aggressive Classifier operates by iteratively updating its model parameters to minimize loss while making predictions. In the context of a drug recommendation system, PAC can be applied to classify drug reviews as positive or negative based on user ratings. Initially, the classifier is trained on labeled data, where it learns to distinguish between positive and negative ratings. During training, if a misclassification occurs, the model updates its parameters aggressively to correct the mistake. However, when correctly classified instances are encountered, the model remains passive and does not update its parameters. This unique behavior allows PAC to adapt quickly to changing data patterns while maintaining

efficiency. In the drug recommendation system, PAC can effectively classify reviews, aiding in the generation of accurate medication recommendations based on user ratings. The system benefits from PAC's ability to handle large datasets and adapt to evolving user preferences, ultimately enhancing the overall performance and reliability of the recommendation process.

Multinomial Naive Bayes: Multinomial Naive Bayes (MNB) is a probabilistic classifier commonly used in text classification tasks, including drug recommendation systems. It operates based on the assumption that features are conditionally independent given the class label. In the context of drug reviews, MNB assigns probabilities to each word's occurrence in a review given the class label (positive or negative rating). These probabilities are estimated from the training data using a multinomial distribution. During classification, MNB calculates the probability of a review belonging to each class using Bayes' theorem and selects the class with the highest probability as the predicted rating for the review. Within drug recommendation systems, Multinomial Naive Bayes demonstrates efficiency in handling extensive text data. It rapidly categorizes reviews as either positive or negative, relying on their associated ratings. This capability significantly enhances the system's accuracy in delivering personalized and precise medication recommendations, aligning with individual preferences and requirements.

3.5 Classifiers:

Various machine learning classification algorithms were employed to develop classifiers for users rating. Passive Aggressive Classifier and Multinomial Naive Bayes were applied to the Word2Vec and manual features model, as these algorithms are better suited for handling sparse matrices compared to tree-based classifiers, which could be computationally intensive. Given the large dataset comprising approximately 210,000 reviews, computational efficiency was a significant concern. Therefore, algorithms were selected based on their ability to reduce training time and provide faster predictions without compromising accuracy.

3.6 Metrics:

Five metrics were examined in order to evaluate the user ratings: accuracy, precision, recall, F1 score, and AUC score. Here, Tp stands for times when the model correctly predicted positive sentiment, Tn for times when the model correctly predicted negative sentiment, Fp for times when the model incorrectly predicted positive sentiment, and Fn for times when the model incorrectly predicted negative sentiment. These evaluations were used to calculate the F1 score, precision, recall, and accuracy.

$$\text{Precision} = Tp/Tp + Fp$$

$$\text{Recall} = Tp/Tp + Fn$$

$$\text{Accuracy} = Tp + Tn/Tp + Tn + Fp + Fn$$

$$\text{F1score} = 2.\ \text{Precision}.\ \text{Recall}/\text{Precision} + \text{Recall}$$

The Area Under Curve score serves as a metric for assessing a classifier's ability to differentiate between classes. It is commonly utilized in the evaluation of the Receiver Operating Characteristic curve. This curve illustrates the relationship between the true positive rate and false positive rate across different thresholds.

3.7 RECOMMENDER SYSTEM FOR MEDICATIONS:

After assessing the metrics, the most effective predictions from the four classifiers were amalgamated to form a consolidated prediction. These combined results were then multiplied by the normalized useful count to generate an overall score for a specific condition. A higher score indicates a more favorable drug. The normalization of the useful count was prompted by the observed distribution, where the range between the minimum and maximum counts is significant, at approximately 1300, with a notable deviation of 36. This normalization was implemented because higher useful counts may not necessarily indicate better drug performance, as increased searches could lead to more reviews irrespective of user ratings. Hence, normalizing the useful count by condition was deemed essential for constructing an efficient recommender system.

IV. WORKING OF DRUG RECOMMENDER SYSTEM:

A prevalent methodology involves incorporating machine learning and user ratings into drug recommendation systems, marking a groundbreaking approach to improving medical decision-making processes. This cutting-edge system harnesses the capabilities of sophisticated algorithms to scrutinize extensive datasets that encompass drug reviews, patient feedback, and medical literature. Employing machine learning techniques, notably collaborative filtering and content-based filtering, the system adeptly discerns patterns and correlations within the data, ultimately generating personalized medication recommendations.

Furthermore, user ratings play a vital part in augmenting the accuracy and relevance of these recommendations of customer. By evaluating user ratings in patient reviews, algorithms can discern the overall satisfaction levels, concerns, and preferences associated with specific medications. This invaluable insight is then integrated into the recommendation model to prioritize medications with favorable user ratings while considering potential side effects highlighted in unfavorable reviews.

The synergy between machine learning empowers the drug recommender system to deliver tailored and evidence-based recommendations that cater to individual patient needs and preferences. By facilitating informed decision-making for both healthcare providers and patients, this system aims to optimize treatment outcomes, improve medication adherence, and enhance the overall quality of healthcare delivery. Through continuous refinement and optimization, the integration of machine learning in drug recommendation systems holds immense promise in revolutionizing the landscape of personalized medicine and patient care.

Table 2. Recommendations for the top four medications for each of the top five medical conditions.

Condition	Drug Name	Score
Acne	Retin-A	0.066334
Acne	Atrain	0.088545
Acne	Magnesium hydroxide	0.088545
Acne	Retin A Micro	0.097399
Birth Control	Mono-Linyah	0.005448
Birth Control	Glides Fe 1.5/30	0.005987
Birth Control	Ortho Micronor	0.006149
Birth Control	Lybrel	0.027766
High BP	Adalat CC	0.303191
High BP	Zestril	0.305851
High BP	Toprol-XL	0.362589
High BP	Labetalol	0.367021
Pain	Neurontin	0.158466
Pain	Nortriptyline	0.171771

Pain	Pamelor	0.231829
Pain	Elavil	0.304513
Depression	Remeron	0.124601
Depression	Sinequan	0.146486
Depression	Provigil	0.240185
Depression	Methylin ER	0.328604

V. RESULT:

Based on the consumer's star rating, each review in this study was categorized as either favorable or unfavorable. Positive ratings were those with five stars or more, and negative ratings were those with a range of one to five stars. Initially, the training data included 111,583 positive ratings and 47,522 negative ratings. To address class imbalances, algorithms were applied, adjusting the minority class to constitute 70% of the majority class examples. As a result, the updated training data comprised 111,583 positive classes and 78,108 negative classes. Four distinct text representation methods – BoW, TF-IDF, Word2Vec, and Manual feature – were employed alongside ten diverse ML algorithms for binary classification. The outcomes were assessed using five different metrics.

VI. DISCUSSION:

While the results obtained from each of the methods are promising, they do not necessarily indicate readiness for real-life applications. There is still room for improvement in the recommender framework. The observed discrepancies between the class metrics that are favorable and unfavorable. suggest the need for balancing the training data appropriately. Additionally, proper hyperparameter optimization is essential for classification algorithms to enhance model accuracy. In the recommendation framework, simply combining the best-predicted results of each method may not suffice. Improved results and comprehension necessitate the implementation of proper ensemble techniques for integrating different predicted outcomes. It is important to note that this paper primarily focuses on demonstrating the methodology for extracting user rating from data and performing classification to develop a recommender system.

VII. CONCLUSION:

In this study, we delve into user ratings of drug reviews to construct a recommender system, leveraging various machine learning classifiers. We employed classifiers were applied on Word2Vec and Manual features. Evaluation was conducted using precision, recall, F1 score, accuracy, and AUC score metrics. Notably, Linear SVC on TF-IDF exhibited superior performance with 95% accuracy. Additionally, we combined the best-predicted rating values from each method and multiplied them by the normalized useful count to derive an overall drug score by condition for the recommender system. Future endeavors will explore different oversampling techniques, n-gram values, and algorithm optimization to further enhance the recommender system's performance.

Table 4 Performance metrics

Classifier	Accuracy	Precision	Recall	F1 Score
Passive Aggressive	0.956033	0.956201	0.956033	0.955916
LightGBM	0.935995	0.937354	0.935995	0.935939
Decision Tree	0.8193392	0.812910	0.813392	0.806700
Cat Boost	0.808851	0.807544	0.808851	0.801866
Multinomial NB	0.666066	0.765140	0.666066	0.610787
Random Forest	0.666066	0.765140	0.666066	0.610787

VIII. REFERENCES:

[1] Jay Prakash Gupta, Ashutosh Singh, A Computer-Based Disease Prediction and Medicine Recommendation System Using Machine Learning Approach, SRM University AP Amravati, Andra Pradesh, India.

[2] D. Vigneswari, N. Komal Kumar, A Drug Recommendation System for Multi-disease in Health Care Using Machine Learning, Advances in Communication and Computational Technology, Singapore, 2021

[3] Wittich CM, Burkle CM, Lanier WL. Medication errors: an overview for clinicians. Mayo Clin Proc. 2014.

[4] GV Lavanya, Praveen KS, Drug recommender system using machine learning for sentiment analysis, International Research Journal of Modernization in Engineering Technology and Science, Bangalore, Karnataka, India

[5] Satvik Garg, Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning, Jaypee University of Information Technology Solan, India.

[6] Drug Recommendation System Using NLP and Machine Learning Approach in Python, 2023 https://wisdomml.in/drug-recommendation-system-using-nlp-and-machine-learning-approach-in-python/

[7] Priyanka V.G, Pushpalatha G, Drug Recommendation System Using Machine Learning, International Journal of Arts, Science and Humanities, 2023