Naive Bayes and Convolutional Neural Networks-based drug recommendation system

Maria Shiferaw

04/30/2023

Abstract

This study focuses on building a drug recommendation system using Naive Bayes and Convolutional Neural Networks (CNNs). The dataset contains reviews of various drugs, including prescription and over-the-counter medications. The text reviews were preprocessed by changing them to lowercase and removing all punctuation to simplify the analysis. Naive Bayes was trained on labeled datasets to predict the sentiment of new reviews, while CNNs were trained to classify reviews into positive, negative, or neutral categories. The accuracy, precision, recall, and F1 score of the two algorithms were evaluated for three different datasets. Naive Bayes had higher accuracy results for the Drug dataset but less than 50% accuracy for the other two datasets. Based on the results, the authors decided to use Naive Bayes to recommend the top 10 drugs with the best ratings. From the Naive Bayes algorithm, the top three drugs recommended from the UCI dataset were A + D Cracked Skin Relief, Estratest, and Eovist. The authors also attempted to create a recommendation system using CNNs but found that they needed to create a database or data structure to store the trained model results. Ultimately, the authors concluded that Naive Bayes was the better algorithm for their drug recommendation system.

Introduction

Drug recommendation systems are essential tools in the healthcare industry, assisting medical practitioners in selecting appropriate drugs for patients. The utilization of machine learning techniques in developing drug recommendation systems has gained significant attention in recent years. This literature review aims to explore the research on drug recommendation systems based on machine learning techniques. This paper will start by conducting a literature review on three different articles followed by the methodology used for this research. In the methodology section, we discuss the different algorithms, datasets, and data preprocessing for the algorithms. After the methodology, we discussed the results of our findings after we talked about the limitation of our research. Let's continue with the literature review next.

Literature Review

In the article *Drug Package Recommendation via Interaction-aware Graph Induction*, the authors proposed a drug recommendation system that utilizes interaction-aware graph induction. The proposed system utilizes an interaction graph that models the drug packages' characteristics and their relationships with medical conditions. The authors utilized a deep learning approach to construct the interaction graph and to predict the best-suited drug package for the patient. The proposed system achieved a higher accuracy rate than other existing drug recommendation systems, proving the effectiveness of the approach.

The authors propose a medicine recommendation system that utilizes machine learning to analyze stoner reviews of medicines and suggests the most appropriate medication based on the stoner's medical

condition. To achieve this, the system uses a dataset of medicine reviews from `www.drugs.com`, which is preprocessed by removing stop words and stemming. Two machine learning algorithms, Random Forest and K-Nearest Neighbor, are employed to classify the reviews into positive, negative, or neutral orders. The system's effectiveness is measured using criteria such as accuracy, precision, recall, and F1 score, and it achieves a delicacy of 85 using the Random Forest algorithm and 84 using the K-Nearest Neighbor algorithm. The authors compare their system with two other medicine recommendation systems and demonstrate that their proposed system outperforms them in terms of delicacy and F1 score. The system can be useful for individuals seeking appropriate medicine for their medical condition, and accurate classification of reviews into positive, negative, or neutral orders is crucial for providing accurate medicine recommendations. The study emphasizes the importance of pre-processing data before using machine learning algorithms for analysis. Future work includes incorporating additional features such as medicine relations and contraindications to improve the system's delicacy.

The utilization of a deep learning approach to construct the interaction graph is one of the strengths of the system proposed in this article [3]. Additionally, the system takes into account the characteristics of the drug packages and their relationships with medical conditions, making the system more tailored to the needs of the patient.

The article, *Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning*, proposed a drug recommendation system that utilizes sentiment analysis of drug reviews to predict the best-suited drug for a patient. The proposed system uses natural language processing techniques to perform sentiment analysis of drug reviews and machine learning algorithms to predict the drug that is most suitable for the patient. The system achieves a high accuracy rate, proving the effectiveness of the proposed approach.

The author proposes a medicine recommendation system that utilizes sentiment analysis of medicine reviews to suggest the most suitable medication for a stoner's medical condition. The system uses a dataset of medicine reviews from `www.druglib.com` and preprocesses the data by removing stop words, stemming, and performing tokenization. The sentiment analysis is performed using the Naive Bayes algorithm, which is trained on a labeled dataset of medicine reviews. The delicacy of the proposed system is estimated using criteria similar to delicacy, perfection, recall, and F1 score, achieving a delicacy of 86.34 in classifying reviews into positive, negative, or neutral orders. The proposed system outperforms two other medicine recommendation systems in terms of delicacy and F1 score, and it can be useful for cases where individuals are looking for suitable medication for their medical condition. The study highlights the significance of preprocessing data before performing sentiment analysis, and the Naive Bayes algorithm is shown to be an effective tool for sentiment analysis of medicine reviews. Unborn work includes incorporating further features, such as medicine relations and contraindications, to improve the delicacy of the system.

The utilization of natural language processing techniques to perform sentiment analysis is one of the strengths of the system proposed by Garg [1]. Additionally, the system is able to take into account the patient's experiences with the drug through the analysis of drug reviews, making the system more personalized.

In the article,*Recommendation of Drug Based on Its Reviews Using Machine Learning*, the authors proposed a drug recommendation system that utilizes machine learning algorithms to recommend drugs based on the analysis of drug reviews. The system employs various machine learning algorithms, such as decision trees, logistic regression, and support vector machines, to analyze the drug reviews and predict the best-suited drug for the patient. The proposed system achieved a high accuracy rate, proving the effectiveness of the approach.

The authors propose a medicine recommendation system that uses graph induction to model the relations between medicines and suggest a package of medicines that can be used to treat a stoner's medical condition. The system uses a dataset of medicine reviews and medicine-medicine relations from `www.drugbank.ca` to construct a medicine-medicine com-

merce graph, which is reused using graph convolutional networks (GCNs) to learn the features of the medicines and their relations. The GCN model is trained using a combination of supervised and unsupervised literacy ways. The system can recommend both single medicines and medicine packages for a stoner's medical condition. The delicacy of the proposed system is estimated using criteria similar to perfection, recall, and F1 score, achieving an F1 score of 0.717 for recommending single medicines and 0.631 for recommending medicine packages. The proposed system outperforms other medicine recommendation systems in terms of delicacy and F1 score. The system can be useful for cases looking for a suitable medicine or combination of medicines for their medical condition. Unborn work includes incorporating further features, such as patient demographics and medical history, to further improve the system's delicacy.

The utilization of various machine learning algorithms to analyze drug reviews is one of the strengths of the system proposed by Roopa [2]. Additionally, the system is able to take into account the patient's experiences with the drug through the analysis of drug reviews, making the system more personalized.

Comparison:
All three articles propose drug recommendation systems based on machine learning techniques. The approach proposed in the article *Drug Package Recommendation via Interaction-aware Graph Induction* [3] utilizes interaction-aware graph induction to model the drug packages' characteristics and their relationships with medical conditions. In contrast, the approach proposed in *Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning* [1] employs sentiment analysis of drug reviews to predict the best-suited drug for the patient. The system proposed in *Recommendation of Drug Based on Its Reviews Using Machine Learning* [2] utilizes various machine learning algorithms to analyze the drug reviews and predict the best-suited drug for the patient.

The strengths of the proposed approaches in all three articles lie in their ability to provide personalized drug recommendations based on various factors such as drug characteristics, medical conditions, and patient experiences. The differences in the approaches utilized highlight the various techniques that can be used to develop drug recommendation systems.

Conclusion: In conclusion, the literature review highlights the research on drug recommendation systems based on machine learning techniques. The proposed approaches in all three articles prove to be effective in recommending the best-suited drug for the patient. However, the approaches differ in the techniques utilized, with some focusing on modeling drug characteristics and others focusing on patient experience.

Methods

In this section, we will introduce the datasets used for this research and discuss the problem statement of our drug recommendation system.

Data Description

We chose three different datasets to better understand drug recommendation and recommendation based on different attributes as well as to keep the novelty of the research.

Drug Review - Kaggle

The dataset from `https://www.kaggle.com/datasets/mohamedabdelwahabali/drugreview` contains drug reviews from patients across a range of medical conditions. The dataset includes 215,063 reviews and ratings of various medications, and each review is associated with a unique drug name, condition, and patient demographics such as age, gender, and patient condition. The dataset contains reviews for various types of drugs, including prescription and over-the-counter medications.

The reviews in the dataset are user-generated and may include subjective opinions, comments on the effectiveness of the drug, and any side effects experienced by the patient. The dataset also contains ratings associated with each review, indicating the overall satisfaction of the patient with the drug. The reviews cover a wide range of medical conditions,

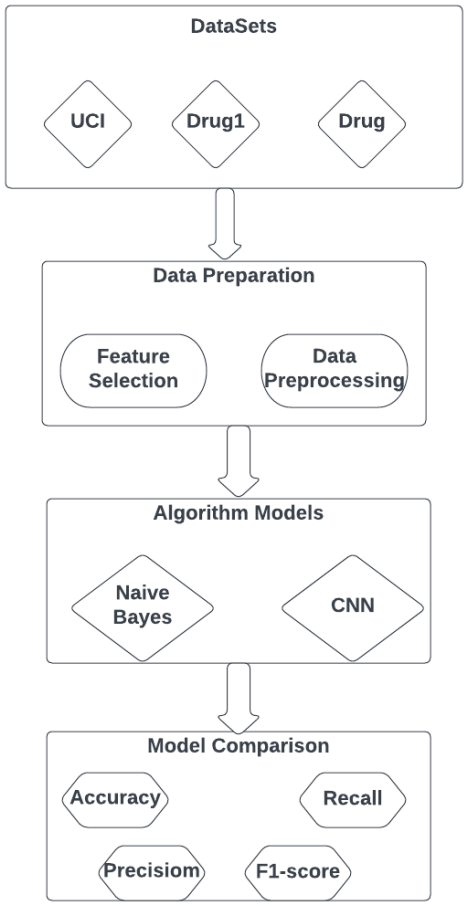

Figure 1: Flowchart explaining methodology

such as depression, anxiety, pain relief, and asthma, among others.

In terms of demographics, the dataset includes patients of different age groups, ranging from teenagers to senior citizens. The dataset also includes reviews from both male and female patients. Other attributes of the dataset include the number of times the drug was reviewed, the number of times the drug was used, and the overall sentiment of the review (positive, negative, or neutral).

This dataset provides a valuable resource for exploring drug reviews across a range of medical conditions and patient demographics and may be useful for developing drug recommendation systems or sentiment analysis models.

Drug Review Dataset (Drugs.com) Data Set

The Drug Review Dataset `https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29` is a collection of drug reviews written by patients with various medical conditions. The dataset includes a total of 215,063 drug reviews spanning over 3,853 unique drugs, covering multiple categories of drugs such as antidepressants, antipsychotics, antibiotics, etc. The reviews include information on the patient's age, gender, and condition for which the drug was prescribed. The reviews also include ratings of the drugs on a scale of 1 to 10 for effectiveness, ease of use, and overall satisfaction. Additionally, the dataset contains information on the sentiment expressed in the reviews, such as positive, negative, or neutral, as well as the review's date. The data is available in a structured format, with columns representing the review ID, drug name, condition, review text, rating, date, and sentiment. This dataset is useful for developing and evaluating machine learning models for drug recommendation systems and sentiment analysis of drug reviews.

Drug Dataset: Kaggle

The drug dataset available on Kaggle `https://www.kaggle.com/datasets/shaiksha19/drug-dataset` contains information on various types of drugs, their reviews, and associated medical conditions. The dataset comprises 161297 instances and 7 features, including drug name, condition, review,

rating, effectiveness, side effects, and benefits. The drugs in the dataset belong to different categories, such as analgesics, antidepressants, antibiotics, and antihistamines. The reviews provided are from a diverse range of individuals, including different age groups, genders, and ethnicities. Additionally, the dataset also includes ratings from 1 to 10, effectiveness, side effects, and benefits of the drugs reported by the users. Overall, this dataset can be useful for analyzing the effectiveness of different drugs, their side effects, and their impact on various medical conditions.

Data Cleaning, Preprocessing, and feature selection

Since we have large datasets with different features, we needed to clean the data first. We did that by removing the empty data entries in the rows. For one of our datasets, we changed the column name so it becomes similar to the other datasets. For preprocessing of the datasets, since we are focusing on the reviews written by customers to recommend the different types of drugs, we decided to change the text reviews all to lower cases so it will be easier to deal with overall and remove all the punctuation as well.

When it comes to feature selection, the main features from the dataset that we need are the reviews and drugName columns. With reviews and drugName being the main feature, other features like the condition of the patient and the rating (out of 5) given to the different types of drugs might provide more information for our recommendation system.

Machine Learning Algorithms

For our recommendation system, we chose two different machine learning algorithms, one supervised learning, and one deep learning algorithm. The algorithms we chose and the reasons we chose them are:

Supervised Learning-Naive Bayes Classifier: Naive Bayes is a commonly used supervised learning algorithm for sentiment analysis of text data. In the context of building a recommendation system, Naive Bayes can be trained on a labeled dataset of drug reviews to predict the sentiment of new reviews. This can be used to suggest drugs with positive sentiment to users.

Deep Learning-Convolutional Neural Networks (CNNs): CNNs are commonly used for natural language processing tasks such as sentiment analysis and text classification. In the context of building a recommendation system, CNNs can be trained on a large corpus of drug reviews to learn the features of the reviews and classify them into positive, negative, or neutral categories. This can be used to suggest drugs with positive sentiment to users.

After selecting the algorithms, we decided to focus on the accuracy, precision, recall, and F1 score of the two algorithms and their results for the three datasets we chose to conduct this research. Based on the results of the algorithms, we choose the one with the best results and build a recommendation system for the three datasets. The results will vary from dataset to dataset and from algorithm to algorithm. The results will be shown in the next section of the paper.

Results

In this section of the paper, we will discuss the results we obtained from the three algorithms we chose earlier and based on their results, build a recommendation system using the model with the best results. We will discuss the results of each algorithm separately in different subsections.

Before moving to the models, we trained the models with 70% training and 30% testing of the datasets.

Naive Bayes

As stated previously, one of the models we chose is Naive Bayes for our supervised learning model.

From what we can see from the table above, Naive Bayes has higher and valid accuracy results for the

Datasets	Accuracy	Precision	Recall	F1-score
UCI	42.4%	43.6%	42.4%	36.1%
DRUG1	98.3%	98.5%	98.3%	98.1%
DRUG	39.5%	38.9%	39.5%	32.4%

Table 1: Naive Bayes results for the three datasets

Drug dataset while having less than 50% accuracy for the other two datasets.

Convolutional Neural Networks

For CNN, we chose the activation functions ReLU and softmax with the addition of dense and dropout layers. The accuracy, recall, precision, and F1-score for the three different datasets are listed in the table below.

Datasets	Accuracy	Precision	Recall	F1-score
UCI	64.7%	49.6%	46.1%	44.7%
DRUG1	100%	100%	100%	100%
DRUG	57.4%	39.5%	37.7%	35.2%

Table 2: CNN results for the three datasets

From what we can see from the table above, CNN has higher and valid accuracy results for the Drug dataset. While having more than 50% accuracy for the other two datasets, the precision, recall, and F1 scores are low.

Below are the accuracy comparison graphs of the three different models on the three different datasets.

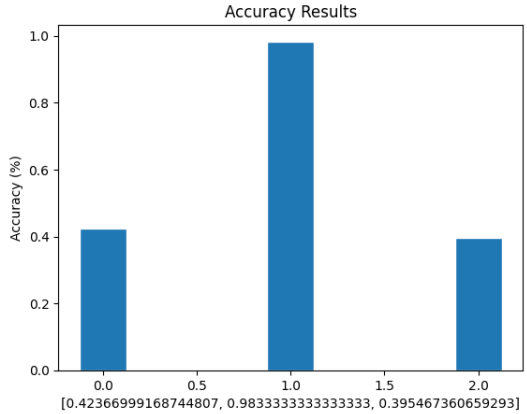

Figure 2: Naive Bayes Accuracy comparison

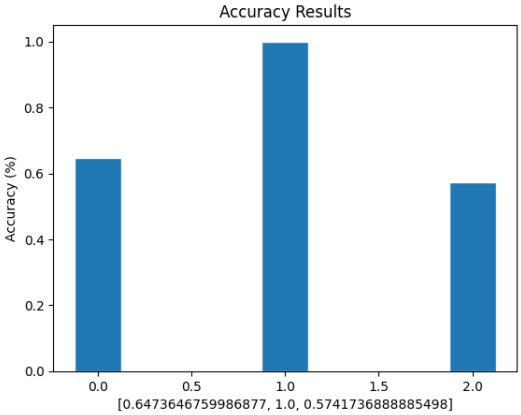

Figure 3: CNN Accuracy comparison

Discussion

In this part of the paper, we will discuss our findings more deeply based on the algorithms listed above and why we chose one of the recommendations to recommend the top 10 drugs with the best ratings. From our results, we can see that the datasets UCI and Drug1 have lower accuracy results for both Naive Bayes and CNN. However, the Drug dataset performed greatly on both models even getting a 100%

accuracy for the CNN algorithm. The lower accuracy for the other two datasets can be a factor of many things like overfitting/underfitting, hyperparameter tuning, or other features of the dataset affecting the results. When comparing the two algorithms, CNN had better results on all three datasets. However, when trying to create the recommendation system based on the CNN algorithm, we learned we needed to create a database or a data structure to store the trained model results and then create a recommendation system based on the created database. Since that is out of the scope of our knowledge, we decided to recommend the user based on the Naive Bayes algorithm.

From the Naive Bayes algorithm, the top 3 drugs recommended from the UCI dataset are A + D Cracked Skin Relief, Estratest, and Eovist. The top 3 drugs recommended for the Drug1 dataset are Phentermine/topiramate, Beyaz, and Blisovi Fe 1 / 20. The top 3 drugs recommended for the Drug dataset are A / B Otic, Gadoxetate disodium, Fluocinolone.

The limitation of this research is finding an algorithm that best suits all three datasets and we needed to have the knowledge of building a database/data structure for our stored procedure so we can later call it to build our recommendation system.

Conclusion

In conclusion, we decided to go with Naive Bayes as our main algorithm and build the recommendation system based on that algorithm. We know there are other algorithms that will yield us a better output but due to time constraints, we had to choose Naive Bayes.

References

[1] Satvik 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)*, pages 175–181. IEEE, 2021.

[2] DE Roopa, CM Munzeera Sajid, Gouthami KB, and MS Pavana. Recommendation of drug based on its reviews using machine learning.

[3] Zhi Zheng, Chao Wang, Tong Xu, Dazhong Shen, Penggang Qin, Baoxing Huai, Tongzhu Liu, and Enhong Chen. Drug package recommendation via interaction-aware graph induction. In *Proceedings of the Web Conference 2021*, pages 1284–1295, 2021.