

DRUG RECOMMENDATION SYSTEM BASED ON SENTIMENT ANALYSIS OF DRUG REVIEWS USING MACHINE LEARNING

O.HARITHA¹

Dr. M. VEERA KUMARI , M-Tech, Ph.D²

M-Tech Student, Dept.of CSE,UNIVERSAL COLLEGE OF ENGINEERING AND TECHNOLOGY,Andhra pradesh, India ¹

Associate Professor,HOD Dept.of CSE,UNIVERSAL COLLEGE OF ENGINEERING AND TECHNOLOGY,Andhra pradesh, India²

oppicharlaharitha@gmail.com , veerakumarimamidi@gmail.com

ABSTRACT — Due to issues such as an absence of qualified medical staff, insufficient hospital infrastructure, and essential drugs and medical equipment, the accessibility of high-quality healthcare resources has been steadily decreasing since their establishment. Consequently, individuals all over the globe are placing a heavy burden on healthcare systems and, in some cases, endangering their own health by attempting to self-medicate. A lot of people are interested in machine learning because of its potential to help with healthcare decision-making by automating and augmenting it. This study proposes a medication recommendation system to assist patients in selecting appropriate medications, with the aim of facilitating the work of healthcare professionals. The system assesses sentiment and makes predictions based on patient medication ratings using a host of text representation algorithms, such as Bag of Words (BoW), TF-IDF, Word2Vec, and proprietary feature extraction methods. A number of classification methods are used to assess the processed textual input and to forecast its emotion. The models are evaluated using common evaluation metrics such as recall, accuracy, precision, F1-score, and AUC score. Combining the Linear Support Vector Classifier (LinearSVC) with TF-IDF vectorisation yields the best performance, according to the experimental results. This strategy outperforms all others that were considered. When used together, these factors provide an accuracy of around 93%. Intelligent drug recommendation systems developed using machine learning-based sentiment analysis have the potential to improve healthcare accessibility and clinical decision-making, as shown here.

Keywords: Issues of Access, Innovative Medicines, Vectorisation, and Emotions.

I. INTRODUCTION

There is a severe scarcity of competent medical professionals, which is adding more strain on healthcare systems worldwide as the number of COVID-19 cases continues to rise at an alarming rate. Because of the already restricted availability of medical experts in rural areas, this problem is much more acute there. It is not feasible to rapidly expand the supply of healthcare workers to satisfy increasing demand since being a fully qualified doctor involves around six to twelve years of schooling and clinical experience. In these types of situations, telemedicine technologies play a crucial role in providing additional healthcare help. Medical mistakes are still a major issue in the field of clinical practice. Even in modern healthcare systems, reports show that prescription mistakes impact a considerable number of individuals annually. Prescriptions are often based on individual

clinical experience, which doesn't always include all medications, illnesses, and patient situations. This is one of the main reasons why this is the case. New medications, research discoveries, and treatment strategies are launched on a daily basis, making it even more difficult to choose the best suitable prescription for a patient. Taking into account symptoms, medical history, and medication interactions becomes more and more complex for doctors when making decisions. In the meantime, user-generated product and medication evaluations have proliferated thanks to the expansion of online marketplaces. Analysing these evaluations with machine learning approaches may help with clinical decision-making and enhance medication recommendation systems by providing useful insights into patients' experiences.

A. Problem Statement:

Drug Review Dataset (Drugs.com) was the dataset used in this endeavour. A copy of it was taken from the UCI Machine Learning Storage. Medicine name, review, medical condition, review date, useful count (a numerical number that indicates how many users considered the review helpful), and overall patient satisfaction (a 10-point rating scale) are the six components that make up this dataset. This dataset is perfect for training and testing machine learning recommendation systems, since it has over 215,063 items in total. Overall, the system is structured to prepare data, classify data, evaluate data, and provide suggestions.

A. Data Cleaning and Visualization

To make sure the data is consistent and of high quality, the data is preprocessed using industry standard methods. Finding and dealing with missing information, deleting duplicate records, and filtering out unnecessary or noisy items are all part of this process. In order to enhance the quality of feature extraction, textual data is additionally sanitised.

The integrity of the dataset is ensured during preparation by meticulously removing duplicate items by checking unique IDs. Furthermore, to deal with records that are lacking values in important columns like the "condition" column, we remove incomplete rows (around 1200 null entries). Following these procedures will guarantee that your dataset is ready for analysis and model training.

B. Challenges in Drug Recommendation:

1. Data Quality and Availability

- Why Datasets used for drug reviews can include inaccurate, missing, or biased information.
- Slang, spelling problems, or unrelated information could be included in reviews.
- Structured clinical data is scarce, which lowers precision.

2. Imbalanced Datasets

- There are thousands of reviews for some medications but relatively few for others.
- It's not uncommon for there to be more positive evaluations than negative ones.
- Model performance and the accuracy of predictions are both impacted by this mismatch.

3. Natural Language Processing (NLP) Complexity

- Review sarcasm and irony
 - Medical jargon unique to a particular field
 - Opinions that differ within a one review
- Necessitates state-of-the-art methods in NLP.

4. Context Awareness

- It is common for models to ignore:
 - Age, gender, and health background
 - How long and what dosage
- Less effective due to lack of personalisation.

C. Proposed Work Aim:

Through sentiment analysis and prediction algorithms, the system hopes to glean useful information from textual reviews and recommend the best treatments for individual medical problems.

Objectives of the Project

The objective is to create and assess several machine learning algorithms, such as Support Vector Machine (SVM), Logistic Regression, and Naive Bayes, for accurate drug classification and prediction.

In order to make better use of feature extraction approaches like TF-IDF and make the model more efficient and provide more accurate suggestions.

In order to tailor drug recommendations to each patient based on their specific symptoms and medical history.

To make sure that critical aspects including efficacy, side effects, and patient ratings are taken into account when medication recommendations are made.

Creating a web-based or GUI that is simple to use and navigate is the goal.

Assisting both patients and healthcare providers in making better drug selection decisions.

D. Objectives:

1) *Data Collection and Preparation*

- Accumulate databases of medication reviews that include ratings, user comments, medication names, and medical issues.
- To prepare the data for analysis by eliminating noise, dealing with missing numbers, and normalising text.

2) *Text Processing using NLP*

- Using Natural Language Processing methods to glean useful information from review texts.
- To tidy up review text by tokenising, removing stop words, stemming/lemmatizing, and so on.

3) *Sentiment Analysis*

- To determine whether user reviews are good, negative, or neutral by analysing them.
- Using sentiment ratings to assess medication efficacy and patient satisfaction.

4) *Feature Extraction*

- Transforming textual data into numerical form may be accomplished using the following methods:
- One such approach is the TF-IDF, or Term Frequency-Inverse Document Frequency.
- The Word Fanny Pack
- To improve the model's performance by picking out relevant details.

5) *Model Development*

The goal is to: Apply and evaluate several ML algorithms, including:

- Regression Optimisation
- SVM, or Reverse Vector Machine
- Bayesian
- options: ○ Random Forest ○ Multilayer Perceptron

6) *Drug Recommendation System*

- Constructing a system that does medication recommendation based on:
 - People's health
 - the findings of the sentiment analysis
 - Drug efficacy and rating systems

E. Overview of the Paper:

This article describes how to construct a smart drug recommendation system using machine learning and Natural Language Processing (NLP). This system might help doctors and patients choose the best medications. This approach analyses patient evaluations and large-scale medical reviews to find out how well pharmaceuticals work, how safe they are, and how satisfied users are with the treatment overall. Emphasising the positive, negative, and neutral aspects of patient attitudes, sentiment analysis clarifies the real-world impacts of drugs. Feature extraction methods, like TF-IDF and Bag of Words, enhance prediction accuracy by numerically representing the textual input. The following ML models are evaluated and utilised to find the best method for medication recommendation: There are several varieties of neural networks, such as logistic regression, naive bayes, random forests, multilayer perceptrons, and support vector machines (SVMs). After the user enters their symptoms or medical difficulties, the system learns from past data and then delivers personalised medication options. Our suggestions are more relevant and trustworthy since we incorporate information

like user reviews, sentiment ratings, and the perceived efficacy of medicine. Everyone from patients to physicians will have no problem navigating the well-designed interface. Standard metrics for evaluating the robustness and functioning of the system include recall, accuracy, precision, and F1-score. This initiative seeks to improve medication recommendation systems by connecting patient experiences with clinical decision-making using data-driven intelligence.

II. LITERATURE SURVEY

Intelligent medication recommendation systems that use Machine Learning and Natural Language Processing (NLP) have been the subject of several investigations. Screening user-submitted pharmaceutical evaluations, recognising sentiment trends, and creating prescription recommendations still relies heavily on human reviewers. Prior work in this field mostly used sentiment analysis and other conventional text mining methods applied to massive pharmaceutical review databases. Word2Vec, Bag of Words, and TF-IDF are examples of algorithms that often extract numerical information from text. Then, we use machine learning algorithms like Logistic Regression, Naive Bayes, and Support Vector Machine (SVM) to classify sentiment and deliver medicine recommendations. Accuracy ratings of about 93% are not uncommon when TF-IDF representation is employed in conjunction with Linear SVM, according to experimental data from these research.

Prediction accuracy and contextual knowledge have recently been enhanced by breakthroughs in deep learning-based methodologies. Improve the quality of recommendations and sentiment classification accuracy using models like Bidirectional Long Short-Term Memory (BiLSTM) networks and others by understanding the sequential and contextual relationships seen in pharmaceutical evaluations.

The integration of organised clinical data with free-form textual evaluations is another patient-centered method that several academics have created. When it comes to understanding semantic meaning and providing more tailored medical options based on patient-specific facts and experiences, these technologies often use complex language models like BERT and Word2Vec.

Experts in both machine learning and domain knowledge, including a patient's medical records and information on drug interactions, have come up with the idea of hybrid recommendation systems. To improve the accuracy and safety of suggestions, many approaches are used, such as deep learning and matrix factorisation.

When faced with complicated and enormous datasets, research into alternative classifiers like Multilayer Perceptron (MLP), Passive Aggressive methods, and LightGBM has shown promising results in terms of efficiency and scalability. These developments show how smart pharmaceutical recommendation systems are becoming better all the time, which leads to better, more tailored healthcare.

III. PROPOSED METHODOLOGY

The suggested system is built to recommend suitable drugs by evaluating patient medication evaluations with the help of machine learning and Natural Language Processing (NLP). To guarantee precise processing, analysis, and suggestion, the whole approach is structured into a number of methodical steps. Efficacious sentiment analysis and dependable medication prediction are made possible by each step of the system, which is built to convert raw textual data into valuable insights. Starting with data preparation, the workflow proceeds to feature extraction, model training, assessment, and, eventually, user-input based medicine recommendation.

1. Data Collection

- Publicly accessible sources are used while compiling drug review databases.
 - In most cases, the dataset will include:
 - medicine brand
 - a health problem
 - the text of a user review
 - A numerical rating

2. Cleaning Up Your Data

For better quality and consistency, raw data is cleaned:

- Eliminate values that are empty or do not exist
 - Clearing the database of duplicate records
 - Changing all capital letters to lowercase
 - Eliminating punctuation and special letters

Section 3: Text Analysis

- NLP methods are used to process the review text:
 - "tokenisation" means breaking text into individual words.
 - Eliminating commonly used but unneeded words
- De-emphasising words by stemming or lemmatising them

4. Extraction of Features

- The process of transforming text into numerical representation involves:
 - The Term Frequency-Inverse Document Frequency (TF-IDF) method.
 - The Bag of Words
- To better grasp textual input, machine learning models benefit from these properties.

5. Analysing Emotions

- There are three ways reviews may be categorised:
 - Positive,
 - Negative, and
 - Neutral.

- To measure how well drugs work and how satisfied users are, researchers utilise sentiment ratings.

6. Developing the Model

- Various algorithms for machine learning are used:
 - Regression Optimisation
 - SVM, or Reverse Vector Machine
 - Bayesian
 - options: o Random Forest o Multilayer Perceptron
- Feature extraction and sentiment labelling are used to train models.

7. Tests and Training of Models

- A training set (e.g., 80%) is created from the dataset.
- The testing set, for instance, 20%
- The data that is not viewed is used to train and assess the models.

8. Assessing Performance

- To evaluate the model's efficacy, we use:
 - a precision
 - being exact
 - Remember
 - my F1-score
- Iteratively, the most effective model is chosen for implementation.

9. The Module for Drug Recommendation

- Medications are suggested by the system depending on:
 - user-provided information (health status, symptoms)
 - Scores for emotions
 - drug evaluations
- By using the trained model, the last suggestion is produced.

*A. Algorithm:***B. TF-IDF Formula (Core Concept)**

$TF\text{-}IDF(w,d) = TF(w,d) \times \log \left(\frac{N}{DF(w)} \right)$
 $TF(w,d) = \text{Number of times } w \text{ appears in } d$
 $IDF(w,d) = \log \left(\frac{N}{DF(w)} \right)$
 Where:

- $TF(w,d)$ = Term frequency of word w in document d
- N = Total number of documents
- $DF(w)$ = Number of documents containing word w

BEGIN**1. LOAD Dataset**

INPUT: Drug reviews dataset (text, label)

2. DATA PREPROCESSING (NLP)

FOR each review in dataset DO

- Convert text to lowercase
- Remove punctuation, numbers, special characters
- Tokenize text (split into words)
- Remove stopwords (e.g., "is", "the", "and")
- Apply stemming or lemmatization
- Store cleaned text

END FOR

3. BUILD VOCABULARY

- Collect all unique words from cleaned dataset

4. COMPUTE TERM FREQUENCY (TF)

FOR each document d in dataset DO

FOR each word w in vocabulary DO

$TF(w,d) = \frac{\text{Number of times } w \text{ appears in } d}{\text{Total words in } d}$

END FOR

END FOR

5. COMPUTE INVERSE DOCUMENT FREQUENCY (IDF)

FOR each word w in vocabulary DO
 $IDF(w) = \log \left(\frac{\text{Total number of documents}}{\text{Number of documents containing } w} \right)$
 END FOR

6. COMPUTE TF-IDF MATRIX

FOR each document d DO

FOR each word w DO

$TFIDF(w,d) = TF(w,d) * IDF(w)$

END FOR

END FOR

7. FEATURE VECTOR CREATION

- Convert each document into TF-IDF vector

8. SPLIT DATA

- Training set and Testing set

9. TRAIN MODEL

- Apply ML algorithm (SVM / Naive Bayes / Logistic Regression)

10. PREDICTION

INPUT: New drug review

- Apply same preprocessing

- Convert into TF-IDF vector

- Predict sentiment (Positive / Negative)

11. RECOMMENDATION

IF sentiment is Positive THEN

Recommend drug

ELSE

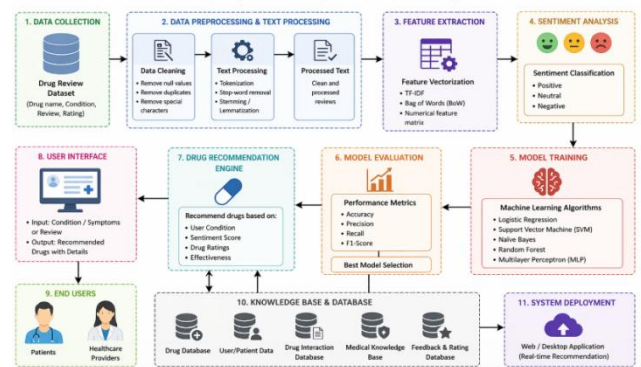
Do not recommend

END IF

END

System Architecture:**System Architecture of Drug Recommendation System**

(Using Sentiment Analysis and Machine Learning)

**IV. RESULTS AND DISCUSSIONS****A. Data Collection and Preprocessing**

Collecting raw data from several sources is what this stage entails.

1) Data Sources

Research tools: questionnaires, experiments, and sensors

Reference materials such as databases, application programming interfaces, and scholarly articles

Information gathered from the web, including: data scraping, social media, and logs

Private information: books, records, and financial details

2) Main Things to Think About

Relevance of data (does it align with your goal)

Can it be analysed with this amount of data?

Information integrity (truthful, exhaustive, uniform)

Problems with ethics and the law (GDPR, consent, privacy)

In most cases, raw data is in a disorganised state. A clean, useable format is achieved by preprocessing.

3) Main Steps**1. Data Cleaning**

- *a) Deal with missing values (deleting, filling, or assuming)*
- *Eliminate imposters*
- *Correct mistakes (such as typos or mismatched formats)*

2. Data Integration

- Bring together information from many sources
- Sort out disagreements (such as between formats or units)

3. Data Transformation

- Make sure all data is standardised
- For example, one-hot encoding may be used to encode categorical variables.
- Scaling features

4. Data Reduction

- Get rid of extraneous elements

- De-emphasising (e.g., principal component analysis)
- Combining

5. Data Discretization (if needed)

- To classify continuous data (such as age groups), one must:

B. Data Augmentation

A medicine recommendation system that uses sentiment analysis of pharmaceutical reviews leverages data augmentation to manage limited or imbalanced datasets, improve model performance, and handle text data.

C. 1. Why Data Augmentation is Needed

- Inadequate or skewed drug review databases (with more favourable than negative evaluations, for example) are possible.
- Contributes to a more accurate NLP model
- Cuts down on overfitting
- Makes models more resistant to language variances

D. 2. Types of Text Data Augmentation**1) 1. Synonym Replacement**

Replace words with similar meanings.

Example:

- Original: "This drug is very effective"
- Augmented: "This medicine is highly effective"

2) 2. Random Insertion

Insert new words randomly.

Example:

- "This drug works well" → "This drug really works well"

3) 3. Random Deletion

Remove some words without losing meaning.

Example:

- “This drug is very effective” → “This drug effective”

4) 4. Random Swap

Swap positions of words.

Example:

- “Drug works very well” → “Very drug works well”

V. CONCLUSION

These days, reviews are something people check out before making any kind of decision, whether it about a restaurant, a product, or even an online company. This work uses sentiment analysis of pharmaceutical reviews and several machine learning methods to construct a smart recommendation system based on this pattern of behaviour. Multiple classification techniques were used in the study. These algorithms included Logistic Regression, Perceptron, Multinomial Naive Bayes, Ridge Classifier, Stochastic Gradient Descent, and Linear Support Vector Classifier (LinearSVC). Some of the feature extraction approaches used to construct these models include Bag of Words (BoW), TF-IDF, Word2Vec, and features that we developed ourselves. Try using boosting and tree-based approaches such as Decision Tree, Random Forest, LightGBM, and Random Forest for even better results with a broad range of feature representations. One common way to evaluate a model's efficacy is by looking at its precision, accuracy, recall, F1-score, and AUC. According to the results, the LinearSVC model that included TF-IDF features had the best accuracy (around 93 percent) throughout the trials. A much lower accuracy of around 78% is achieved by Decision Tree classifiers trained with Word2Vec features. Before building the final recommendation system, all feature-method combinations are considered, including the best emotion ratings. Nearly everyone uses LightGBM in conjunction with Word2Vec; nearly everyone uses LinearSVC with TF-IDF; nearly everyone uses Random Forest with user-defined parameters; and only a tiny fraction uses Perceptron with Bag of Words. Incorporating these expected emotion

values with normalised "usefulCount" values yields a drug effectiveness score for every medical reason. This aggregated score helps provide pharmaceutical choices that are more accurate and condition-specific.

A) Scope for Future Work

In order to better manage class imbalance and enhance model performance, future work for this project will include investigating and evaluating various oversampling strategies. Also, trying out different n-gram configurations may improve feature representation by capturing more contextual information from medication review language. To further increase the recommendation system's accuracy, resilience, and efficiency, hyperparameter tweaking and algorithm optimisation may be used to machine learning. All things considered, these updates should make the system more dependable for use in actual healthcare settings and raise the bar for the quality of predictions made generally.

VI REFERENCES

- [1] Ministry of Health and Family Welfare, “Telemedicine Practice Guidelines,” Government of India. Available: <https://www.mohfw.gov.in/pdf/Telemedicine.pdf>
- [2] Wittich, C. M., Burkle, C. M., & Lanier, W. L. “Medication errors: an overview for clinicians.” Mayo Clinic Proceedings, vol. 89, no. 8, pp. 1116–1125, 2014.
- [3] Chen, M. R., & Wang, H. F. “The reason and prevention of hospital medication errors.” Practical Journal of Clinical Medicine, vol. 4, 2013.
- [4] UCI Machine Learning Repository, “Drug Review Dataset (Drugs.com).” Available:

<https://archive.ics.uci.edu/ml/datasets/Drug%2BReview%2BDataset%2B%2528Drugs.com%2529>

- [5] Fox, S., & Duggan, M. "Health Online 2013." Pew Internet Research Report. Available: <http://pewinternet.org/Reports/2013/Health-online.aspx>
- [6] Bartlett, J. G., Dowell, S. F., Mandell, L. A., et al. "Practice guidelines for the management of community-acquired pneumonia in adults." *Clinical Infectious Diseases*, vol. 31, no. 2, pp. 347–382, 2000.
- [7] Fox, S., & Duggan, M. "Health Online 2013." Pew Research Internet Project Report, 2012.
- [8] Tekade, T. N., & Emmanuel, M. "Probabilistic aspect mining approach for interpretation and evaluation of drug reviews." in *Proc. IEEE SCOPES*, 2016, pp. 1471–1476.
- [9] Doulaverakis, C., Nikolaidis, G., Kleontas, A., et al. "GalenOWL: Ontology-based drug recommendations discovery." *Journal of Biomedical Semantics*, vol. 3, 2012.
- [10] Sun, L., Liu, C., Guo, C., Xiong, H., & Xie, Y. "Data-driven automatic treatment regimen development and recommendation." in *Proc. ACM KDD*, 2016, pp. 1865–1874.
- [11] Goel, V., Gupta, A. K., & Kumar, N. "Sentiment analysis of multilingual Twitter data using NLP." in *IEEE CSNT*, 2018, pp. 208–212.
- [12] Shimada, K., Takada, H., Mitsuyama, S., et al. "Drug-recommendation system for patients with infectious diseases." *AMIA Annual Symposium Proceedings*, 2005.
- [13] Bao, Y., & Jiang, X. "An intelligent medicine recommender system framework." in *IEEE ICIEA*, 2016, pp. 1383–1388.
- [14] Zhang, Y., Zhang, D., Hassan, M., Alamri, A., & Peng, L. "CADRE: Cloud-assisted drug recommendation service for online pharmacies." *Mobile Networks and Applications*, vol. 20, pp. 348–355, 2014.
- [15] Li, J., Xu, H., He, X., Deng, J., & Sun, X. "Tweet modeling with LSTM recurrent neural networks for hashtag recommendation." in *IEEE IJCNN*, 2016, pp. 1570–1577.

FIRST AUTHORS:

O.HARITHA pursuing her M.Tech in Computer Science And Engineering in Universal College Of Engineering And Technology.

SECOND AUTHOR:

Dr. M. VEERA KUMARI, M.Tech, Ph.D and B.Tech received her all degree's in Computer Science And Engineering. She is currently working as Associate Professor and HOD of CSE in Universal College Of Engineering And Technology.

