DRUG RECOMMENDATION SYSTEM USING MACHINE LEARNING BASED ON TF-IDF FEATURE EXTRACTION PROCESS

Rajaiah M[1,a)] ,Vijaya Kumar C[2,b)], Subramanyam N[3,c)] and Chandra Kanth P[4,d)]

[1]*Professor &HOD, Department of Computer Science Engineering, Audisankara College of Engineering and Technology, Gudur,India*

[2,3]*PG Scholar, Department of Computer Science Engineering, Audisankara College of Engineering and Technology, Gudur,India*

[4]*Associate Professor,Department of Computer Science Engineering, Audisankara College of Engineering and Technology, Gudur,India*

[a)] rajagopal1402@gmail.com
[b)] Corresponding author: vijay.mba.vsu@gmail.com
[c)]subramanyam.naguru@gmail.com
[d)] chandrakanthc4u@gmail.com

ABSTRACT

Since the coronavirus emerged, there has been a rise in the inaccessibility of genuine clinical resources, such as a shortfall of experts and healthcare workers, a lack of appropriate equipment and medications, and so on. Many people have died as a result of the medical community's crisis.

Due to the lack of availability, individuals began taking medication on their own without proper consultation, worsening their health condition. Machine learning has lately shown to be beneficial in a variety of applications, resulting in an increase in new automation projects. The purpose of this project is to create a pharmaceutical recommendation system that relieves doctors of a significant amount of work. We developed a drug recommendation system that predicts sentiment based on patient ratings using the TF-IDF vectorization technique for this project.

1. INTRODUCTION

The countries are experiencing a doctor crisis as the number of coronavirus infections rises at an alarming pace, particularly in rural locations where there are fewer experts than cities. Obtaining the requisite skills to be a doctor could take from around 6 to 12 years. As a result, the number of doctors cannot be expanded in the required time. During this difficult time, automated assistance should be used as much as possible [1] and also medical accidents are all too prevalent. Prescription mistakes injure around 200 thousand persons in China and 100 million in the United States each year. Over 40%

of medical professionals make blunders while prescribing since they create the answer based on their inadequate understanding. [2][3]. Suggesting the best prescription is absolutely essential for patients needing specialists who are well-versed [6]. Every day, fresh research is published, and new drugs and diagnostics are made accessible to clinicians. As a result, doctors are having a more difficult time determining which medication or medicines to prescribe to a patient according to the indications and their clinical history. Product evaluations have become a vital and integral part in acquiring things all over the world, thanks to the exponential rise of the internet and the web-based commercial industry. Before making a purchasing choice, people all over the world have developed the practice of reading the reviews and browsing websites. While most prior study has focused on analyzing E-Commerce expectations and ideas, medical care or therapeutic treatments have received less attention. There is a sudden increase in the number of people who are concerned about their health and seek a diagnosis online. In accordance a 2013 Pew American Research Center research [5], around 60% of respondents explore the internet for health-related subjects, with roughly 35% seeking for health-related diagnoses. A drug recommendation system help both doctors and patients in aiding their diseases with ease and precision. A recommendation system is a sort of system that recommends an item to a user using their benefits and requirement accordingly. Using consumer surveys we examine their attitudes and provide recommendations based on their specific requirements. Sentiment analysis and extraction of features are used by the medicine recommender system to provide prescription recommendations based on patient feedback. Sentiment analysis is used to recognize and classify subjective data such as attitude and opinions of an individual [7]. Feature engineering, in contrast, is an act of enhancing the performance of the older systems.

2. LITERATURE SURVEY

2.1.” Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. Lancet. 2000;356(9237):1255–9. “

We characterize an unfavorable drug response as "an appreciably harmful or unpleasant reaction caused by an intervention related to the use of a medicinal product, which predicts hazard from future administration and necessitates prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product." The WHO's Adverse Reaction

Terminology, which will ultimately become part of the International Classification of Diseases, is being used to track these reactions. There are six different types of adverse medication responses (with mnemonics): dose-related (Augmented), non-dose-related (Bizarre), dose-related and time-related (Chronic), time-related (Delayed), withdrawal (End of use), and therapy failure (Failure). Timing, illness pattern, investigation results, and rechallenge can all aid in determining the cause of a suspected adverse medication response. Management entails, if possible, drug withdrawal and specific treatment of the drug's side effects. Drug responses that are harmful should be reported if they are suspected. Methods of surveillance can identify emotions and build associations.

2.2 "Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Côté S, et al. Large-scale prediction and testing of drug activity on side-effect targets. Nature. 2012;486(7403): 361–7".

Using just empirical methodologies, it is difficult to uncover inadvertent 'off-targets' that predict negative pharmaceutical reactions. Drugs can affect on a wide range of protein targets, some of which are irrelevant according to standard molecular metrics, and hundreds of proteins have been linked to negative outcomes. The action of 656 marketed drugs on 73 unforeseen "side-effect" targets is forecasted using a computer model. Approximately half of the predictions were validated, either via the use of proprietary datasets not previously used by the method or through fresh experimental testing. These new off-targets have affinities ranging from 1 nM to 30 nM. We created an association score to identify novel off-targets that described side effects better than any known drug target. The prediction that the synthetic oestrogenchlorotrianisene's stomach discomfort side effect was mediated by its recently found suppression of the enzyme cyclooxygenase-1 was one of these novel connections. In whole-blood platelet aggregation studies, the therapeutic importance of this inhibition was established. This strategy has the potential to be widely employed in drug development to decrease toxicological hazards.

2.3 "Ernst FR, Grizzle AJ. Drug-related morbidity and mortality: updating the cost-of-illness model. J Am Pharm Assoc (1996). 2001;41(2):192–9."

When people take drugs, they might experience a wide range of side effects.

Pharmacotherapeutic measures assist patients in the majority of cases; nonetheless, severe events ranging from modest side effects to death might occur. A drug-related difficulty occurs when a medication's anticipated therapeutic impact is not achieved (DRP). 1 Following the initial pharmacological therapy, a patient may develop one or more DRPs. Researchers discovered that the costs of DRPs exceed the cost of initial drug therapy; in other words, the overall cost of medication-related morbidity and mortality exceeds the cost of the pharmaceuticals. 2,3 DRPs are becoming increasingly commonly recognized as a major and urgent medical concern that is entirely avoidable.

2.4 “Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. BMC Bioinformatics. 2011;12(1):1.”

Adverse medication responses, sometimes known as side effects, have grown into a serious public health. It's one of the most common reasons for medication development failure, as well as drug withdrawal once it's been released on the market. As a result, in silico prediction of likely side effects is of great importance early in the drug development process, before reaching the clinical stages, in order to optimize this long and expensive process and create new efficient and safe medicines for patients.

3. PROPOSED SYSTEM

3.1 IMPLAMENTATION OF TF-IDF FEAUTER EXTRACTION

A numerical metric called term frequency–inverse document frequency (TF-IDF) quantifies how essential a word is in a collection or corpus. In data collection, text analytics, and data for a sample search, it's widely employed as a weighting factor. The TF-IDF value grows in proportion to the number of times a comment occurs inside the text, and is balanced by the number of times the phrase appears in a document, explaining why certain terms occurs more frequently than others. The TF-IDF approach the most used term-weighting strategies nowadays. According to a 2015 survey, TF-IDF is used in 83 percent of digital library message recommendation systems.

Term Frequency – Let’s take a case where we have such a corpus of English textual information and desire to sort them by relevance to the question, "the brown cow." To begin, delete papers that do not include all three terms "the," "brown," and "cow," However, there are still a lot of pages left. We can distinguish them even further by noting how many times each term appears in each text. The term frequency is the number of instances an even in a document. Adjustments are required regularly when the length of documents fluctuates substantially. Hans Peter Luhn created the first form of term weighting (1957).

TF (word) = n/ the number of terms in the document ------ (1)

where n = number of times word appears in document

Inverse Document Frequency – Because the term "the" is so common, term frequency will tend to incorrectly emphasise document that use the word "the" more frequently, while failing to give enough weight to the more meaningful terms "brown" and "cow". Unlike the less-common word "brown" and "cow," the term "the" is not a good keyword for distinguishing relevant and irrelevant documents and terms. As a result, an inverse document frequency factor is included, which reduces the weight of term that appear frequently in the document set and increases the weight of terms that appear rarely.

IDF (word) =log (number of documents / number of documents containing the word) ------ (2)

*TF-IDF (word) = TF (word) * IDF (word) ------ (3)*

3.2 IMPLAMENTATION OF MACHINE LEARNIG ALGORITHMS

MULTILAYER PERCEPTRONS

One sort of feed forward artificial neural network is the multilayer perceptron (MLP) (ANN). The term MLP has two meanings: it may referring to any feed forward ANN and it can also refer to networks made up of multiple layers of perceptrons (with threshold activation).When there is only one hidden layer, multilayer perceptrons are termed as "vanilla" neural networks.

The input layer, the hidden layer, and the output layer are the three node layers that make up an MLP. Each node has a neuron with just a nonlinear activation function, which isolates the input nodes. Backpropagation, a supervised learning method, is used by MLP.MLP varies from linear perceptrons in that it has many layers and non-linear activation. It can differentiate data that isn't separable in a linear fashion.

If each neuron in a recurrent neural network has a linear activation function that converts synaptic weights to output, then any hidden layers may be compressed to a two-layer input-output model using linear algebra. Nonlinear activation functions in certain MLP neurons were designed to mimic the frequency of excitability (or firing) in actual neurons.

As one of the possible solutions to sigmoid numerical issues, the consists in the fact unit (ReLU) is being used more routinely in current deep learning break throughs.

The first is a -1 to 1 hyperbolic tangent, whereas the second is a logistic function with a similar form but a 0 to 1 range. The output of the node (neuron), as well as the weighted total of the input connections, are shown below. The Activation functions which can be utilised as an alternative. Radial basis functions are highly specialised activation functions (used in radial basis networks, another sort of supervised neural network model).

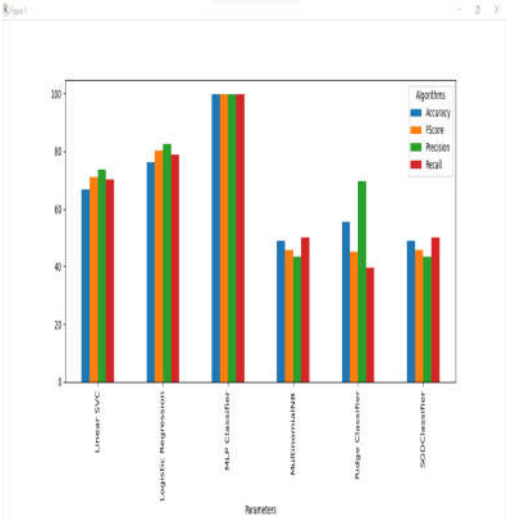


Fig 3.2.1: Accuracy comparison between different machine learning algorithm using TF-IDF

We created the modules listed below to help us carry out this project.

1) Upload Drug Review Dataset: Using this module, we upload the drug review dataset to the application.

2) Read and Pre-process Dataset: Using this module, we read all of the reviews, drug names, and ratings from the dataset and create a features array.

3) TF-IDF Features Extraction: A features array is fed into the TF-IDF algorithm, which calculates the average frequency of each word and then replaces it with the frequency value to form a vector. If the word does not appear in the sentence, the value 0 is used. All reviews will be used as input features for the machine learning algorithm, with RATINGS and Drug Name serving as class labels.

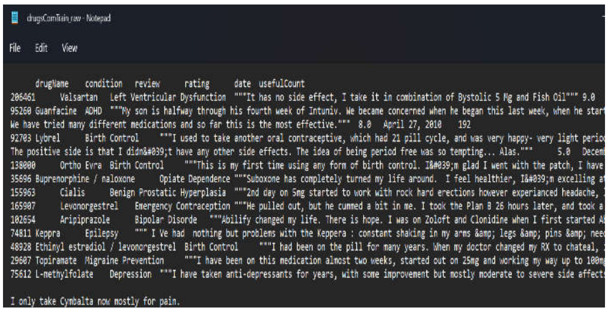
4) Train Machine Learning Algorithms: Using this module, we input TF-IDF features to all machine learning algorithms and then train a model, which is then applied to test data to calculate the algorithm's prediction accuracy.

5) Comparison Graph: We use this module to create an accuracy graph for each algorithm.

6) Recommend Drug from Test Data: We use this module to upload disease name test data, and ML predicts drug name and ratings.

4. DATASET

To implement this project, we used DRUGREVIEW dataset from UCI machine learning website and below is the dataset screen shot.

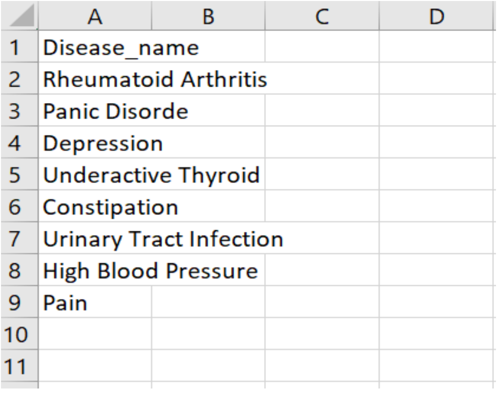


```
      drugName    condition   review      rating      date  usefulCount
206461      Valsartan   Left Ventricular Dysfunction  """It has no side effect, I take it in combination of Bystolic 5 Mg and Fish Oil""" 9.0
95260 Guanfacine  ADHD  """My son is halfway through his fourth week of Intuniv. We became concerned when he began this last week, when he start
We have tried many different medications and so far this is the most effective."""  8.0   April 27, 2010    192
92703 Lybrel      Birth Control     """I used to take another oral contraceptive, which had 21 pill cycle, and was very happy- very light perio
The positive side is that I didn&#039;t have any other side effects. The idea of being period free was so tempting... Alas."""      5.0   Decem
138000      Ortho Evra  Birth Control     """This is my first time using any form of birth control. I&#039;m glad I went with the patch, I have
35696 Buprenorphine / naloxone      Opiate Dependence """Suboxone has completely turned my life around.  I feel healthier, I&#039;m excelling a
155963      Cialis      Benign Prostatic Hyperplasia  """2nd day on 5mg started to work with rock hard erections however experianced headache, 
165907      Levonorgestrel    Emergency Contraception """He pulled out, but he cummed a bit in me. I took the Plan B 26 hours later, and took a
102654      Aripiprazole      Bipolar Disorde   """Abilify changed my life. There is hope. I was on Zoloft and Clonidine when I first started A
74811 Keppra      Epilepsy    """ I Ve had  nothing but problems with the Keppera : constant shaking in my arms &amp; legs &amp; pins &amp; nee
48928 Ethinyl estradiol / levonorgestrel  Birth Control     """I had been on the pill for many years. When my doctor changed my RX to chateal, 
29607 Topiramate  Migraine Prevention     """I have been on this medication almost two weeks, started out on 25mg and working my way up to 100m
75612 L-methylfolate    Depression  """I have taken anti-depressants for years, with some improvement but mostly moderate to severe side affect

I only take Cymbalta now mostly for pain.
```

Fig 4.1: Dataset used for training the model

In above screen first row represents dataset column names such as drug name, condition, review and rating and remaining rows contains dataset values and we will used above REVIEWS and RATINGS to trained machine learning models. Below is the test data which contains only disease name and machine learning will predict Drug name and ratings.

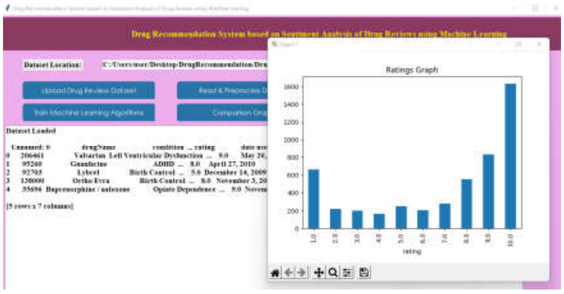


	A	B	C	D
1	Disease_name			
2	Rheumatoid Arthritis			
3	Panic Disorde			
4	Depression			
5	Underactive Thyroid			
6	Constipation			
7	Urinary Tract Infection			
8	High Blood Pressure			
9	Pain			
10				
11				

Fig 4.2: Dataset used to test the model

In above test data we have only disease name and machine learning will predict ratings and drug names.

5. RESULTS AND DISCUSSIONS


Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning
Dataset Location:
C:/Users/user/Desktop/DrugRecommendation/Dru
Upload Drug Review Dataset
Read & Preprocess D
Train Machine Learning Algorithms
Comparison Gra
Ratings Graph
0
200
400
600
800
1000
1200
1400
1600
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
rating

```
Dataset Loaded

  Unnamed: 0            drugName              condition ... rating       date use
0     206461           Valsartan  Left Ventricular Dysfunction ...   9.0      May 20,
1      95260          Guanfacine                  ADHD ...   8.0   April 27, 2010
2      92703              Lybrel         Birth Control ...   5.0  December 14, 2009
3     138000          Ortho Evra         Birth Control ...   8.0  November 3, 20
4      35696  Buprenorphine / naloxone  Opiate Dependence ...   9.0  Novem

[5 rows x 7 columns]
```

Fig 5.1: Uploading dataset

In above graph we can see dataset loaded and in graph x-axis represents ratings and y-axis represents total number of records which got that rating.

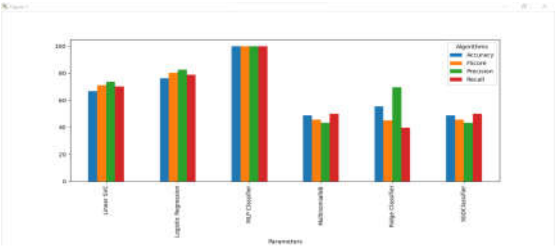
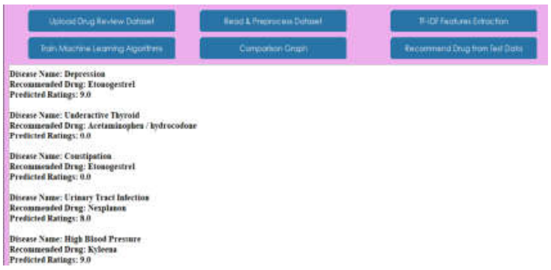


Fig 5.2: Comparison graph between machine learning algorithms

In the following graph, the x-axis shows the algorithm employed, while the y-axis represents accuracy, precision, recall, and FSCORE, with each distinct colour bar representing a different statistic, and we can see that MLP performed well.


Upload Drug Review Dataset
Read & Preprocess Dataset
TF-IDF Features Extraction
Train Machine Learning Algorithms
Comparison Graph
Recommend Drug from Test Data

```
Disease Name: Depression
Recommended Drug: Etonogestrel
Predicted Ratings: 9.0

Disease Name: Underactive Thyroid
Recommended Drug: Acetaminophen / hydrocodone
Predicted Ratings: 0.0

Disease Name: Constipation
Recommended Drug: Etonogestrel
Predicted Ratings: 0.0

Disease Name: Urinary Tract Infection
Recommended Drug: Nexplanon
Predicted Ratings: 8.0

Disease Name: High Blood Pressure
Recommended Drug: Kyleena
Predicted Ratings: 9.0
```

Fig 5.3: Resultant predicted drug output

In above screen for each disease name application has predicted recommended drug name and ratings.

6. CONCLUSION

The development of technology opened a door for development in intelligent recommendation systems.AI Recommendation systems in medical field are helping patients in process of healing their diseases. We use review driven recommendation systems for online shopping in our daily lives. Motivated by this, in this research we used machine learning algorithms such as logistic regression, linear support vector machine, Ridge Classifier, Multinominal Naive Bayes, SGD Classifier, Multilayer perceptron. Using machine learning methods and the TF-IDF algorithm we developed a model that takes the disease names and recommend most appropriate drugs for that disease. Out of all algorithms used TF-IDF is giving more accuracy with multilayer perceptron algorithm. Future scope involves inclusion of more diseases and the drugs related to the disease and optimising the algorithms for better performance of the recommendation system.

REFERENCES

[1]Telemedicine, https://www.mohfw.gov.in/pdf/Telemedicine.pdf

[2] Wittich CM, Burkle CM, Lanier WL. Medication errors: an overview for clinicians. Mayo Clin Proc. 2014 Aug;89(8):1116-25.

[3] CHEN, M. R., & WANG, H. F. (2013). The reason and prevention of hospital medication errors. Practical Journal of Clinical Medicine, 4.

[4] Drug Review Dataset, https://archive.ics.uci.edu/ml/datasets/Drug% 2BReview%2BDataset%2B%2528Drugs.com%2529#

[5] Fox, Susannah, and Maeve Duggan. ”Health online 2013. 2013.” URL: http://pewinternet.org/Reports/2013/Health-online.aspx

[6] Bartlett JG, Dowell SF, Mandell LA, File TM Jr, Musher DM, Fine MJ. Practice guidelines for the management of community-acquired pneumonia in adults. Infectious Diseases Society of America. Clin Infect Dis. 2000 Aug;31(2):347-82. doi: 10.1086/313954. Epub 2000 Sep 7. PMID: 10987697; PMCID: PMC7109923.

[7] Fox, Susannah & Duggan, Maeve. (2012). Health Online 2013. Pew Research Internet Project Report.

[8] T. N. Tekade and M. Emmanuel, ”Probabilistic aspect mining approach for interpretation and evaluation of drug reviews,” 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, 2016, pp. 1471-1476, doi: 10.1109/SCOPES.2016.7955684.

[9] Doulaverakis, C., Nikolaidis, G., Kleontas, A. et al. GalenOWL: Ontology-based drug recommendations discovery. J Biomed Semant 3, 14 (2012). https://doi.org/10.1186/2041-1480-3-14

[10] Leilei Sun, Chuanren Liu, Chonghui Guo, Hui Xiong, and Yanming Xie. 2016. Data-driven Automatic Treatment Regimen Development and Recommendation. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1865–1874. DOI:https://doi.org/10.1145/2939672.2939866

[11] V. Goel, A. K. Gupta and N. Kumar, "Sentiment Analysis of Multilingual Twitter Data using Natural Language Processing," 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2018, pp. 208-212, doi: 10.1109/CSNT.2018.8820254.

[12] Shimada K, Takada H, Mitsuyama S, et al. Drug-recommendation system for patients with infectious diseases. AMIA Annu Symp Proc. 2005;2005:1112.

[13] Y. Bao and X. Jiang, "An intelligent medicine recommender system framework," 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), Hefei, 2016, pp. 1383-1388, doi: 10.1109/ICIEA.2016.7603801.