



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

IJRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.423

Drug Recommendation System Based on Sentiment Analysis of Drug Reviews Using Machine Learning

Mr. S. Ismail Saheeb¹, Shaik Sana Nahida², Mara Charitha³, D. Sai Sandhya⁴, Godyala Pravallika⁵,
Maligaveli Ravindra Reddy⁶

UG Student, Dept. of E.C.E., GATES Institute of Technology, Gooty, Anantapur (Dist.), Andhra Pradesh, India^{2,3,4,5,6}

Assistant Professor, Dept. of E.C.E., GATES Institute of Technology, Gooty, Anantapur (Dist.), Andhra Pradesh,
India¹

ABSTRACT: The lack of legitimate clinical resources, such as specialists and healthcare workers, appropriate equipment, and medicines, etc., is at its highest level since the coronavirus emerged. Numerous people die as a result of the distress experienced by the medical community as a whole. People started taking medications on their own without consulting a doctor because there was none available, which made the health condition worse than usual. Recent years have seen an increase in innovative automation-related work and applications for machine learning.

The goal of this paper is to present a drug recommendation system that has the potential to significantly reduce the heap of specialists. Using various vectorization processes like Bow, TF-IDF, Word2Vec, and Manual Feature Analysis, we build a medicine recommendation system that uses patient reviews to predict sentiment and can help recommend the best drug for a given disease using various classification algorithms. The precision, recall, f1score, accuracy, and AUC score were used to evaluate the predicted sentiments. With an accuracy of 93%, the classifier Linear SVC created with TF-IDF vectorization outperforms all other models.

KEYWORDS: Drug, Recommender System, Machine Learning, NLP, Smote, Bow, TF-IDF, Word2Vec, Sentiment analysis.

I. INTRODUCTION

With the number of coronavirus cases growing exponentially, the nations are facing a shortage of doctors, particularly in rural areas where the quantity of specialists is less compared to urban areas. A doctor takes roughly 6 to 12 years to procure the necessary qualifications. Thus, the number of doctors can't be expanded quickly in a short time frame. A Telemedicine framework ought to be energized as far as possible in this difficult time [1]. Clinical blunders are very regular nowadays.

Over 200 thousand individuals in China and 100 thousand in the USA are affected every year because of prescription mistakes. Over 40% medicine, specialists make mistakes while prescribing since specialists compose the solution as referenced by their knowledge, which is very restricted [2][3]. Choosing the top level medication is significant for patients who need specialists that know wide-based information about microscopic organisms, antibacterial medications, and patients [6].

Every day a new study comes up with accompanying more drugs, tests, accessible for clinical staff every day. Accordingly, it turns out to be progressively challenging for doctors to choose which treatment or medications to give to a patient based on indications, past clinical history. With the exponential development of the web and the web-based business industry, item reviews have become an imperative and integral factor for acquiring items worldwide. Individuals worldwide become adjusted to analyze reviews and websites first before settling on a choice to buy a thing.

While most of past exploration zeroed in on rating expectation and proposals on the E-Commerce field, the territory of medical care or clinical therapies has been infrequently taken care of. There has been an expansion in the number of individuals worried about their well-being and finding a diagnosis online.

II. LITERATURE SURVEY

[1] Wittich CM, Burkle CM, Lanier WL. Medication errors: an overview for clinicians. *Mayo Clin Proc.* 2014 Aug;89(8):1116-25. Medication error is an important cause of patient morbidity and mortality, yet it can be a confusing and underappreciated concept. This article provides a review for practicing physicians that focuses on medication error (1) terminology and definitions, (2) incidence, (3) risk factors, (4) avoidance strategies, and (5) disclosure and legal consequences.

[2] CHEN, M. R., & WANG, H. F. (2013). The reason and prevention of hospital medication errors. *Practical Journal of Clinical Medicine*, Poor prescribing is probably the most common cause of preventable medication errors in hospitals, and many of these events involve junior doctors who have recently graduated. Prescribing is a complex skill that depends on a sound knowledge of medicines, an understanding of the principles of clinical pharmacology, the ability to make judgments concerning risks and benefits, and ideally experience. It is not surprising that errors occur.

[3] 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. Lower respiratory tract infections are the major cause of death in the world and the major cause of death due to infectious diseases in the United States. Recent advances in the field include the identification of new pathogens (*Chlamydia pneumoniae* and hantavirus), new methods of microbial detection (PCR), and new antimicrobial agents (macrolides, b-lactam agents, fluoroquinolones, oxazolidinones, and streptogramins).

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

Thirty-five percent of U.S. adults say that at one time or another they have gone online specifically to try to figure out what medical condition they or someone else might have. These findings come from a national survey by the Pew Research Center's Internet & American Life Project. Throughout this report, we call those who searched for answers on the internet "online diagnoses."

When asked if the information found online led them to think they needed the attention of a medical professional, 46% of online diagnoses say that was the case. Thirty-eight percent of online diagnoses say it was something they could take care of at home and 11% say it was both or in-between. When we asked respondents about the accuracy of their initial diagnosis, they reported: 41% of online diagnoses say a medical professional confirmed their diagnosis.

III. EXISTING SYSTEM

To simplify, assume this feature matrix is binary: a non-zero value means the app has that feature. You also represent the user in the same feature space. Some of the user-related features could be explicitly provided by the user. For example, a user selects "Entertainment apps" in their profile. Other features can be implicit, based on the apps they have previously installed. For example, the user installed another app published by Science R Us.

The model should recommend items relevant to this user. To do so, you must first pick a similarity metric (for example, dot product). Then, you must set up the system to score each candidate item according to this similarity metric. Note that the recommendations are specific to this user, as the model did not use any information about other users.

IV. PROPOSED SYSTEM

The dataset used in this research is Drug Review Dataset (Drugs.com) taken from the UCI ML repository [4]. This dataset contains six attributes, name of drug used (text), review (text) of a patient, condition (text) of a patient, useful count (numerical) which suggest the number of individuals who found the review helpful, date (date) of review entry, and a 10-star patient rating (numerical) determining overall patient contentment.



It contains a total of 215063 instances. Shows the proposed model used to build a medicine recommender system. It contains four stages, specifically, Data preparation, classification, evaluation, and Recommendation. Jim Fulton completes a matrix package, which is later modified by Jim Hugunin to create Numeric, also known as Numerical Python extensions or NumPy. Hugunin, a PhD student at MIT, joined the Corporation for National Research Initiatives (CNRI) to work on JPython in 1997, leaving the maintainer position to Paul Dubois of Lawrence Livermore National Laboratory (LLNL). David Ascher, Konrad Hinsen, and Travis Oliphant were among the early members.

BLOCK DIAGRAM

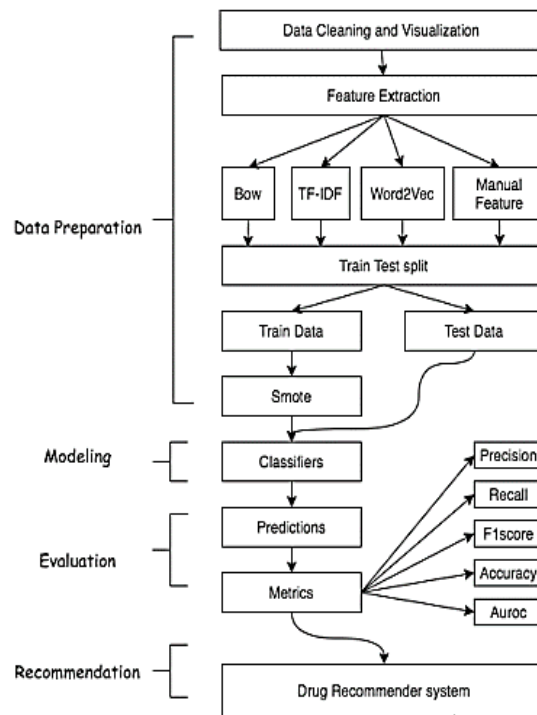


Figure 1. Flowchart of the proposed model

Applied standard Data preparation techniques like checking null values, duplicate rows, removing unnecessary values, and text from rows in this research. Subsequently, removed all 1200 null values rows in the conditions column. We make sure that a unique id should be unique to remove duplicacy.

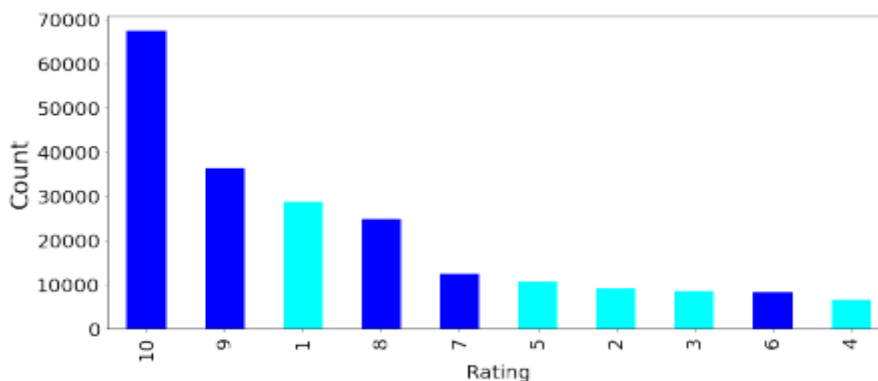


Figure 2. Bar plot of count of rating values versus 10 rating number



After text preprocessing, a proper set up of the data required to build classifiers for sentiment analysis. Machine learning algorithms can't work with text straightforwardly; it should be changed over into numerical format. In particular, vectors of numbers. A well-known and straightforward strategy for feature extraction with text information used in this research is the bag of words (Bow) [16], TF-IDF [17], Word2Vec [18]. Also used some feature engineering techniques to extract features manually from the review column to create another model called manual feature aside from Bow, TF-IDF, and Word2Vec.

TF-IDF: TF-IDF [17] is a popular weighting strategy in which words are offered with weight not count. The principle was to give low importance to the terms that often appear in the dataset, which implies TF-IDF estimates relevance, not a recurrence. Term frequency (TF) can be called the likelihood of locating a word in a document.

After the Train Test split, only the training data has undergone a synthetic minority over-sampling technique (Smote) [22] to prevent the class imbalance problem. Smote is an oversampling technique that synthesized new data from existing data. Smote generates the new minority class data by linear interpolation of randomly selected minority instance 'a' in combination with its k nearest neighbor instance 'b' in the feature space. Table II shows the total distribution of data on final dataset i.e. after data cleaning.

The predicted sentiment were measured using five metrics, namely, precision (Prec), recall (Rec), f1 score (F1), accuracy (Acc.) and AUC score [23]. Let the letter be: T_p = True positive or occurrences where model predicted the positive sentiment truly, T_n = True negative or occurrences where model predicted the negative class truly, F_p = False positive or occurrences where model predicted the positive class falsely, F_n = False negative or occurrences where model predicted the negative class falsely, Precision, recall, accuracy, and f1 score

V. EXPERIMENTAL RESULT

One of the recent diseases affecting humans is the corona virus. Because there aren't enough medical professionals and medical systems to treat these diseases, patients will take medicines at their own risk, which could cause serious side effects or even death.

To address the aforementioned problem, the author of this paper suggests a sentiment- and machine learning-based drug recommendation system. This system will accept the disease names of patients, then recommend a medication and simultaneously display a sentiment rating based on user feedback. If the predicted rating is high, the patient is more likely to take the prescribed medication.

The term frequency – inverse document frequency (TF-IDF), BAG of WORDS, and WORVEC feature extraction algorithms have all been utilized by the author of this paper. Machine learning algorithms like Logistic Regression, Linear SVC, Ridge classifier, Nave Bayes, Multilayer Perceptron classifier, and SGD classifier will all make use of these features. We are combining the preceding algorithm with the TF-IDF features extraction algorithm due to its superior performance.



Figure 3: Precision recall and FS score

VI. ADVANTAGES AND APPLICATIONS

Machine learning has seen use cases ranging from predicting customer behavior to forming the operating system for self-driving cars. When it comes to advantages, machine learning can help enterprises understand their customers at a deeper level. By collecting customer data and correlating it with behaviors over time, machine learning algorithms can learn associations and help teams tailor product development and marketing initiatives to customer demand.

Some companies use machine learning as a primary driver in their business models. Uber, for example, uses algorithms to match drivers with riders. Google uses machine learning to surface the ride advertisements in searches. But machine learning comes with disadvantages. First and foremost, it can be expensive. Machine learning projects are typically driven by data scientists, who command high salaries. These projects also require software infrastructure that can be expensive.

VII. CONCLUSION AND FUTURE SCOPE

Reviews are becoming an integral part of our daily lives; whether go for shopping, purchase something online or go to some restaurant, we first check the reviews to make the right decisions. Motivated by this, in this research sentiment analysis of drug reviews was studied to build a recommender system using different types of machine learning classifiers, such as Logistic Regression, Perceptron, Multinomial Naive Bayes, Ridge classifier, Stochastic gradient descent, Linear SVC, applied on Bow, TF-IDF, and classifiers such as Decision Tree, Random Forest, Lgbm, and Cat boost were applied on Word2Vec and Manual features method. We evaluated them using five different metrics, precision, recall, f1score, accuracy, and AUC score, which reveal that the Linear SVC on TF-IDF outperforms all other models with 93% accuracy.

On the other hand, the Decision tree classifier on Word2Vec showed the worst performance by achieving only 78% accuracy. We added best-predicted emotion values from each method, Perceptron on Bow (91%), Linear SVC on TF-IDF (93%), LGBM on Word2Vec (91%), Random Forest on manual features (88%), and multiply them by the normalized useful Count to get the overall score of the drug by condition to build a recommender system. Future work involves comparison of different oversampling techniques, using different values of n-grams, and optimization of algorithms to improve the performance of the recommender 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. Galen OWL: 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.
- [14] Zhang, Yin & Zhang, Dafang & Hassan, Mohammad & Alamri, Atif & Peng, Limei. (2014). CADRE: Cloud-Assisted Drug Recommendation Service for Online Pharmacies. Mobile Networks and Applications. 20. 348-355. 10.1007/s11036-014-0537-4.
- [15] J. Li, H. Xu, X. He, J. Deng and X. Sun, "Tweet modeling with LSTM recurrent neural networks for hashtag recommendation," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 1570-1577, doi: 10.1109/IJCNN.2016.7727385.
- [16] Zhang, Yin & Jin, Rong & Zhou, Zhi-Hua. (2010). Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics. 1. 43-52. 10.1007/s13042-010-0001-0.
- [17] J. Ramos et al., "Using tf-idf to determine word relevance in document queries," in Proceedings of the first instructional conference on machine learning, vol. 242, pp. 133-142, Piscataway, NJ, 2003.
- [18] Yoav Goldberg and Omer Levy. Word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method, 2014; arXiv: 1402.3722.
- [19] Danushka Bollegala, Takanori Maehara and Kenichi Kawarabayashi. Unsupervised Cross-Domain Word Representation Learning, 2015; arXiv: 1505.07184.
- [20] Textblob, <https://textblob.readthedocs.io/en/dev/>.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details