INTERNATIONAL JOURNAL OF COMPUTATIONAL MATHEMATICAL IDEAS
ISSN:0974-8652/http://sijcmi.weebly.com//VOL14-NO2-PP 1001-1009(2023)

Predication of Drugs in Sentiment Analysis Using Machine Learning Techniques

Velagapudi Sreenivas[1] , Valluri Sai Prasanna[2] ,Valluri Shalini[2], Ponugoti Thirumala Sai Sriram[2], Surapaneni Venkata Sai Krishna[2]

[1] Associate Professor, Department of Computer Science and Engineering, Dhanekula Institute of Engineering and Technology, Vijayawada,A.P,India

[2] B.Tech Students, Department of Computer Science and Engineering, Dhanekula Institute of Engineering and Technology, Vijayawada,A.P,India

Abstract- Since corona virus has shown up, inaccessibility of legitimate clinical resources is at its peak, like the shortage of specialists and healthcare workers, lack of proper equipment and medicines etc. The entire medical fraternity is in distress, which results in numerous individual's demise. Due to unavailability, individuals started taking medication independently without appropriate consultation, making the health condition worse than usual. As of late, machine learning has been valuable in numerous applications, and there is an increase in innovative work for automation. This paper intends to present a drug recommender system that can drastically reduce specialist's heap. In this research, we build a medicine recommendation system that uses patient reviews to predict the sentiment using various vectorization processes like Bow, TF-IDF, Word2Vec, and Manual Feature Analysis, which can help recommend the top drug for a given disease by different classification algorithms. The predicted sentiments were evaluated by precision, recall, f1score, accuracy, and AUC score. The results show that classifier Linear SVC using TF-IDF vectorization outperforms all other models with 93% accuracy.

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

I. INTRODUCTION

With the number of corona virus 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 arevery 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.

As demonstrated in a Pew American Research center survey directed in 2013 [5], roughly 60% of grown-ups searched online for health-related subjects, and around 35% of users looked for diagnosing health conditions on the web. A medication recommender framework is truly vital with the goal that it can assist specialists and help patients to build their knowledge of drugs on specific health conditions. A recommender framework is a customary system that proposes an item to the user, dependent on their advantage and necessity. These frameworks employ the customers' surveys to break down their sentiment and suggest a recommendation for their exact need.

In the drug recommender system, medicine is offered on a specific condition dependent on patient reviews using sentiment analysis and feature engineering. Sentiment analysis is a progression of strategies, methods, and tools for distinguishing and extracting emotional data, such as opinion and attitudes, from language [7]. On the other hand, Featuring engineering is the process of making more features from the existing ones; it improves the performance of models.

This examination work separated into five segments: Introduction area which provides a short insight concerning the need of this research, Related works segment gives a concise insight regarding the previous examinations on this area of study, Methodology part includes the methods adopted in this research, The Result segment evaluates applied model results using various metrics, the Discussion section contains limitations of the framework, and lastly, the conclusion section.

II. LITERATURE REVIEW

With a sharp increment in AI advancement, there has been an exertion in applying machine learning and deep learning strategies to recommender frameworks. These days, recommender frameworks are very regular in the travel industry, e-commerce, restaurant, and so forth. Unfortunately, there are a limited number of studies available in the field of drug proposal framework utilizing sentiment analysis on the grounds that the medication reviews are substantially more intricate to analyze as it incorporates clinical wordings like infection names, reactions, a synthetic names that used in the production of the drug [8]. The study [9] presents Galen OWL, a semantic-empowered online framework, to help specialists discover details on the medications. The paper depicts a framework that suggests drugs for a patient based on the patient's infection, sensitivities, and drug interactions. For empowering Galen OWL, clinical data and terminology first converted to ontological terms utilizing worldwide standards, such as ICD-10 and UNII, and then correctly combined with the clinical information. Leilei Sun [10] examined large scale treatment records to locate the best treatment prescription for patients. The idea was to use an efficient semantic clustering algorithm estimating the similarities between treatment records. Likewise, the author created a framework to assess the adequacy of the suggested treatment. This structure can prescribe the best treatment regimens to new patients as per their demographic locations and medical complications. An Electronic Medical Record (EMR) of patients gathered from numerous clinics for testing.

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. A medication error is any error that occurs at any point in the medication use process. It has been estimated by the Institute of Medicine that medication errors cause 1 of 131 outpatient and 1 of 854 inpatient deaths. Medication factors (eg, similar sounding names, low therapeutic index), patient factors (eg, poor renal or hepatic function, impaired cognition, poly pharmacy), and health care professional factors (eg, use of abbreviations in prescriptions and other communications, cognitive biases) can precipitate medication errors. Consequences faced by physicians after medication errors can include loss of patient trust, civil actions, criminal charges, and medical board discipline. Methods to prevent medication errors from occurring (eg, use of information technology, better drug labeling, and medication reconciliation) have been used with varying success. When an error is discovered, patients expect disclosure that is timely, given in person, and accompanied with an apology and communication of efforts to prevent future errors. Learning more about medication errors may enhance health care professionals' ability to provide safe care to their patients.

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.The challenge of being a prescriber is probably greater now than ever before. Medical education has changed radically in the last 20 years, reflecting concerns about an overburdened curriculum and lack of focus on social sciences. In the UK, these changes have resulted in less teaching in clinical pharmacology and practical prescribing as guaranteed features of undergraduate training and assessment.There has been growing concern, not least from students, that medical school training is not sufficient to prepare them for the pressures of becoming prescribers. Similar concerns are being expressed in other countries. While irrefutable evidence that these changes are related to medication errors identified in practice, there is circumstantial evidence that this is so.Systems analysis of errors suggests that knowledge and training are relevant factors in causation and that focused education improves prescribing performance. We believe that there is already sufficient evidence to support a careful review of how students are trained to become prescribers and how these skills are fostered in the postgraduate years. We provide a list of guiding principles on which training might be based.Medical education has changed greatly in recent years, often for the good. However, it is a matter of regret that specific courses in clinical pharmacology and therapeutics, the discipline that underpins safe and effective prescribing, have been lost. There seems to have been a prevailing view from some that this area of learning will 'take care of itself' as students are exposed to the clinical environment.This has clearly proved to be false. We believe that learning in this area needs to be carefully planned and enthusiastically led for students to achieve the greatest benefit. Teaching and training of prescribers form only part of the approach to protecting patients from medication errors. Support from other colleagues (for example, clinical pharmacists) will be vital, along with the spread of electronic prescribing with decision support, but we believe that it will ultimately be the knowledge and instincts of prescribers that will be their most important protection against irrational and unsafe use of medicines.

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 pneumonia and hantavirus), new methods of microbial detection (PCR), and new antimicrobial agents (macrolides, b-lactam agents, fluoroquinolones, oxazolidinones, and streptogramins).Despite extensive studies, there are few conditions in medicine that are so controversial in terms of management. Guidelines for management were published in 1993 by the American Thoracic Society [1], the British Thoracic Society [2], and the Canadian Infectious Disease Society [3], as well as the Infectious Diseases Society of America (IDSA) in 1998 [4]. The present guidelines represent revised recommendations of the IDSA.Compared with previous guidelines, these guidelines are intended to reflect updated information, provide more extensive recommendations in selected areas, and indicate an evolution of opinion. These therapeutic guidelines are restricted to community-acquired pneumonia (CAP) in immunocompetent adults.Recommendations are given alphabetical ranking to reflect their strength and a Roman numeral ranking to reflect the quality of supporting evidence (table 1). This is customary for quality standards from the IDSA [5]. It should be acknowledged that no set of standards can be constructed to deal with the multitude of variables that influence decisions regarding site of care, diagnostic evaluation, and selection of antibiotics. Thus, these standards should not supplant good clinical judgments.

III.PROPOSED METHODOLOGY

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. Fig. 1 shows the proposed model used to build a medicine recommender system. It contains four stages, specifically, Data preparation, classification, evaluation, and Recommendation.

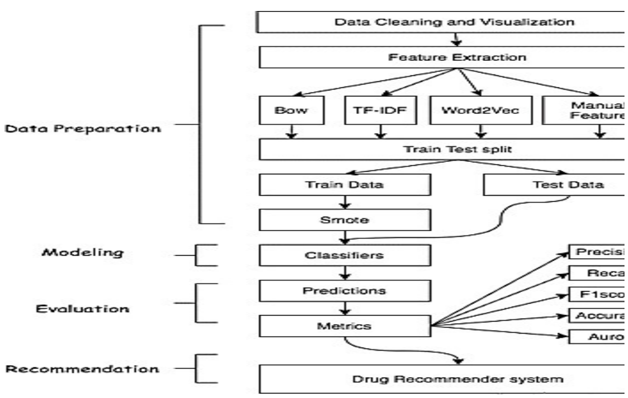


Fig. 1. Flowchart of the proposed model

Data Cleaning and Visualizations

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, as shown in Fig. 2. We make sure that a unique id should be unique to remove duplicacy.

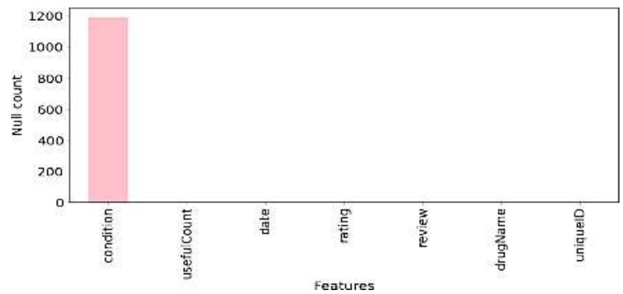


Fig. 2. Bar plot of the number of null values versus attributes

Fig. 3 shows the top 20 conditions that have a maximum number of drugs available. One thing to notice in this figure is that there are two green-colored columns, which shows the conditions that have no meaning. The removal of all these sorts of conditions from final dataset makes the total row count equals to 212141.

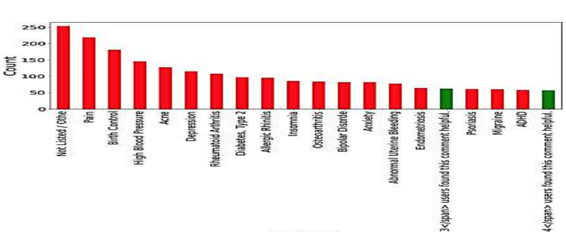


Fig. 3. Bar plot of Top 20 conditions that has a maximum number of drugs available

Fig. 4 shows the visualization of value counts of the 10-star rating system. The rating beneath or equivalent to five featured with cyan tone otherwise blue tone. The vast majority pick four qualities; 10, 9, 1, 8, and 10 are more than twice the same number. It shows that the positive level is higher than the negative, and people's responses are polar. The condition and drug column were joined with review text because the condition and medication words also have predictive power.

Before proceeding to the feature extraction part, it is critical to clean up the review text before vectorization. This process is also known as text

$$idf(t,d) = log(\frac{N}{count(d\epsilon D : t\epsilon d)}) \quad (2)$$

preprocessing. We first cleaned the reviews after removing HTML tags, punctuations, quotes, URLs, etc. The cleaned reviews were lowercased to avoid duplication, and tokenization was performed for converting the texts into small pieces called tokens. Additionally, stop words

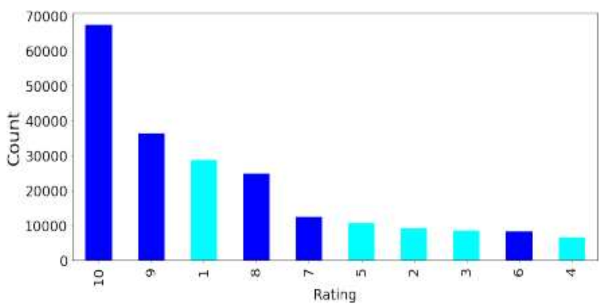


Fig. 4. Bar plot of count of rating values versus 10 rating number

Feature Extraction

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

$$tf(t,d) = log(1 + freq(t,d)) \quad (1)$$

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.

Inverse document frequency (IDF) is the opposite of the number of times a specific term showed up in the whole corpus. It catches how a specific term is document specific.

TF-IDF is the multiplication of TF with IDF, suggesting how vital and relevant a word is in the document

Manual Features:

Feature engineering is a popular concept which helps to increase the accuracy of the model. We used fifteen features, which include use full count, the condition column which is label encoded using label encoder function from Scikit library, day, month, year features were developed from date column using

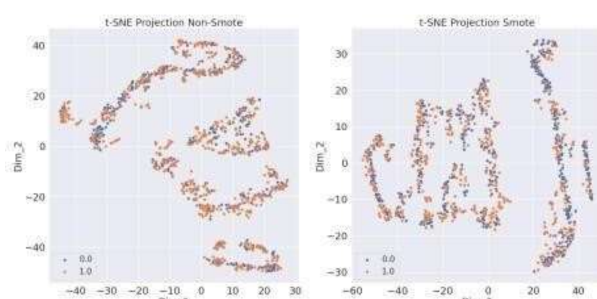
$$tfidf(t,d,D) = tf(t,d).idf(t,D) \quad (3)$$

Date Time function using pandas.

Text blob toolkit [20] was used to extract the cleaned and uncleaned reviews polarity and added as features

$$F1score = 2.\frac{Precision.Recall}{Precision + Recall} \quad (7)$$

along with a total of 8 features generated from each of the text reviews as shown in Table I.



Train Test Split We created four datasets using Bow, TF-IDF, Word2Vec, and manual features. These four datasets were split into 75% of training and 25% of testing. While splitting the data, we set an equal random state to ensure the same set of random

$$Precision = \frac{T_p}{T_p + F_p} \quad (4)$$

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \quad (6)$$

ISSN:0974-8652/http://sijcmi.weebly.com//VOL14-NO2-PP 1001-1009(2023)

numbers generated for the train test split of all four generated datasets.

Fig. 6. T-SNE subplot before and after Smote using 1000 training samples

Metrics

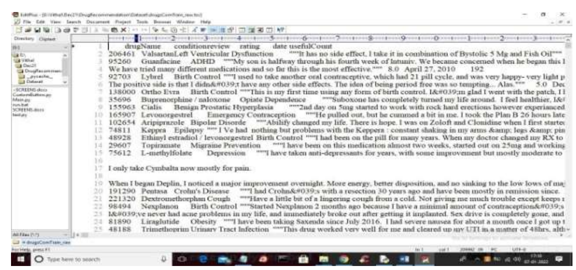
The predicted sentiment were measured using five metrics, namely, precision (Prec), recall (Rec), f1score (F1), accuracy (Acc.) and AUC score [23]. Let the letter be: Tp = True positive or occurrences where model predicted the positive sentiment truly, Tn = True negative or occurrences where model predicted the negative class truly, Fp = False positive or occurrences where model predicted the positive class falsely, Fn = False negative or occurrences where model predicted the negative class falsely, Precision, recall, accuracy, and f1score shown in equations given below,

$$Recall = \frac{T_p}{T_p + F_n} \quad (5)$$

Area under curve (Auc) score helps distinguish a classifier's capacity to compare classes and utilized as a review of the region operating curve (roc) curve. Roc curve visualizes the relationship between true positive rate (Tpr) and false positive rate (Fpr) across various thresholds.

IV. RESULTS & DISCUSSIONS

Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning. Now-a-days new diseases are attacking human world and corona virus is such disease and this diseases require lots of medical systems and medical human experts and due to growing disease medical experts and systems are not sufficient and patients will take medicines on their risk which can cause serious death or serious damage to patient body.To overcome from above problem author of this paper introducing sentiment and machine learning based drug recommendation system which will accept disease names from patient and then recommend DRUG and simultaneously display SENTIMENT rating based on reviews given by old users based on their experience. If predicted rating is high then patient can trust and took recommended drug.In propose paper author has used various features extraction algorithms such as TF-IDF (term frequency – inverse document frequency), BAG of WORDS and WORVEC and this extracted features will be applied on various machine learning algorithm such as Logistic Regression, Linear SVC, Ridge classifier, Naïve Bayes, Multilayer Perceptron classifier, SGD classifier and many more. Among all algorithms TF-IDF is giving better performance so we are using TF-IDF features extraction algorithm with above mention algorithm.To implement this project author has used DRUGREVIEW dataset from UCI machine learning website and below is the dataset screen shots.

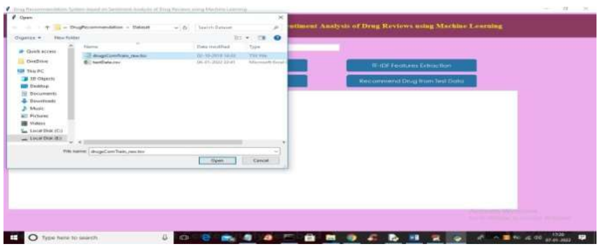


In above screen first row represents dataset column names such as drug name, condition, review and rating and remaining rows contains dataset values and we will used above REVIEWS and RATINGS to trained machine learning models. Below is the test data which contains only disease name and machine learning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

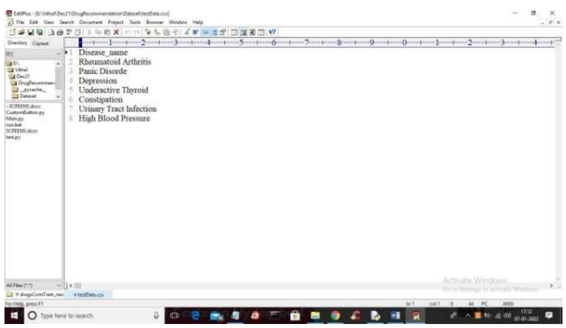
In above test data we have only disease name and machine learning will predict ratings and

drug names. To implement this project we have designed following modules

1)Upload Drug Review Dataset: using this module we will upload dataset to application



2)Read & Pre-process Dataset: using this module we will read all reviews, drug name and ratings from dataset and form a features



array.

3)TF-IDF Features Extraction: features array will be input to TF-IDF algorithm which will find average frequency of each word and then replace that word with frequency value and form a vector. If word not appears in sentence then 0 will be put. All reviews will be consider as input features to machine learning algorithm and RATINGS and Drug Name will be consider as class label.

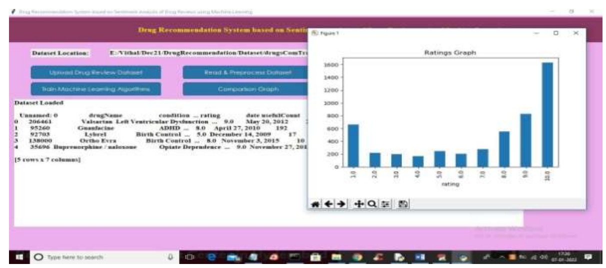
4)Train Machine Learning Algorithms: using this module we will input TF-IDF features to all machine learning algorithms and then trained a model and this model will be applied on test data to calculate prediction accuracy of the algorithm.

5) Comparison Graph: using this module we will plot accuracy graph of each algorithm

6) Recommend Drug from Test Data: using this module we will upload disease name test

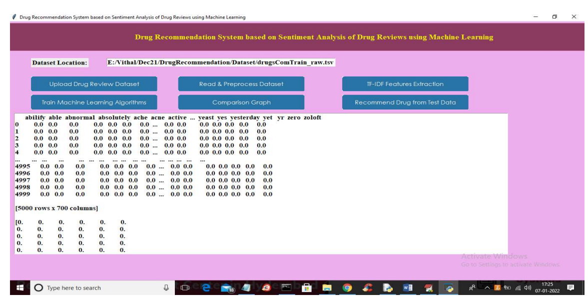
data and then ML will predict drug name and ratings.

To run project double click on 'run.bat' file to get



below screen

In above screen click on 'Upload Drug Review Dataset' button to upload dataset to application and to



get below screen

In above screen selecting and uploading DRUG dataset and then click on 'Open' button to load dataset and to get below screen

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. Now close above graph and then click on 'Read & Preprocess Dataset' button to read all dataset values and then preprocess to remove stop words and special symbols and then form a features array.

In above graph all reviews converted to TF-IDF vector where first row represents review WORDS and remaining columns will contains that word average frequency and if word not appear in review then 0 will put. Now scroll down above screen to view some non-zero frequency values

In above screen for each algorithm we calculate accuracy, precision, recall and FSCORE and in all algorithms MLP has got high performance and now click on 'Comparison Graph' button to get below graph In above screen for each disease name application has predicted recommended drug name andratings

V. CONCLUSION

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/.