IJCSPUB.ORG ISSN : 2250-1770
INTERNATIONAL JOURNAL OF CURRENT SCIENCE (IJCSPUB)
IJCS PUBLICATION (IJCSPUB.ORG) An International Open Access, Peer-reviewed, Refereed Journal

DRUG RECOMMENDATION SYSTEM USING MACHINE LEARNING BASED ON TF-IDFFEATURE EXTRACTION PROCESS - REVIEW

-

K.SOWMYA
Department of Master of Computer Science
Miracle Educational Society Group of Institutions
Vizianagram– 535216 (AP) India

SARAGADAM SRIDHAR
Department of Master of Computer Science
Miracle Educational Society Group of Institutions
Vizianagram– 535216 (AP) India

Abstract

since the outbreak of the coronavirus, there has been an increase in the in accessibility of actual clinical resources, such as a shortage of experts and healthcare professionals, a lack of proper equipment and medications, and so on. As a result of the medical community's problem, many individuals have perished.Individuals began taking medication on their own without sufficient consultation due to a lack of availability, exacerbating their health condition. Machine learning has recently proven to be useful in a wide range of applications,leading in an increase in new automation initiatives. The goal of this project is to develop a pharmaceutical recommendation system that will save doctors a lot oftime.For this research,we used the TF-IDF vectorization technique to createa medicine recommendation system that predicts sentiment based on patient evaluations.

Keywords— (TF-IDF) **term frequency-inverse document frequency**

Introduction

With the quantity of Covid cases developing dramatically, the countries are confronting a lack of specialists,especially in provincial regions where the amount of experts is less contrasted with metropolitan regions. A specialist takes around 6 to 12 years to get the important capabilities. In this way, the quantity of specialists can't be extended rapidly in abriefperiodoftime.A telemedicine structure should best imulated quite far in this troublesome time [1]. Clinical botches are exceptionally standard these days. More than 200 thousand people in China and 100 thousand in the USA are impacted consistently in light ofsolution botches. More than 40% medication, experts commit errors while endorsing sinceexperts create the arrangement as referred to by their insight, which is extremely confined[2][3]. Picking the toplevel drug is huge for patients who need experts that know wide-baseddataaboutinfinitesimallivingbeings,antibacterialmeds,andpatients[6].Consistentlyanother review thinks of going with additional medications, tests, open for clinical staffconsistently. As needs be, it ends up being logically trying for specialists to pick whichtreatment or prescriptions to provide for a patient in view of signs, past clinical history. With the outstanding improvement of the web and the online business industry, thing surveys have turned into an objective and indispensable variable for gaining things around the world.People overall become acclimated to investigate surveys and sites first prior to making adecision to purchase a thing. While the majority of past investigation focused in on rating assumption and recommendations on the E-Commerce field,the domain of clinical consideration or clinical treatments has been rarely dealt with. There has been an extension inthe quantity of people stressed over their prosperity and finding a determination on the web.As shown in a Pew American Research place review coordinated in 2013 [5], generally 60%of adults looked online for wellbeing related subjects, and around 35% of clients searchedfor diagnosing medical issue on th eweb.A prescription recommender structure is really imperative with the objective that it can help trained professionals and assist patients with building the ir insight into drugs on unambiguous medical issue. Arecommenderstructureisa standard framework that proposes a thing to the client, subject for their potential benefit andneed. These structures utilize the clients' studies to separate their feeling and propose asuggestion for their definite need. In the medication recommender framework, medication is presented on a particular condition reliant upon patient surveys utilizing feeling examinationand component designing.Feeling examination is a movement of techniques, strategies, and instruments for recognizing and extricating close to home information, like assessment andmentalities, from language [7]. Then again, Featuring designing is the most common way ofmaking additional elements from the current ones; it works on the exhibition of models. This assessment work isolated into five fragments:Introduction region which gives a short knowledge concerning the need of this exploration, Related works portion gives a compact understanding in regards to the past assessments on this area of study, Methodology part in corporates the techniques embracedin this examination, The Result fragment assesses applied model outcome

Proposed Method

Distinct machine-learning classification algorithms were used to build a classifier topredict the sentiment. After assessing the metrics, all four best-predicted results were picked and joined together to produce the combined prediction.The merged results were then multiplied with normalized useful count to generate an overall score of drug of a particularcondition. The higher the score, the better is the drug. The purpose behind is that the more medications individuals search for,the more individuals read the survey regardless of their review is positive or negative,which makes the useful count high.

ADVANTAGES:

- Accuracy is Veryhigh.
- We selected machine learning classification algorithms only that reduces the training time and give faster predictions.utilizing different measurements, the Discussion segment contains

TF-IDF algorithm

TF-IDF (Term Frequency - Inverse Document Frequency) is a handy algorithm that uses the frequency of words to determine how relevant those words are to a given document. It's a relatively simple but intuitive approach to weighting words, allowing it to act as a great jumping off point for a variety of tasks.

TF-IDF can be broken down into two parts *TF (term frequency)* and *IDF (inverse document frequency).*

TF (term frequency)

Term frequency works by looking at the frequency of a *particular term* you are concerned with relative to the document. There are multiple measures, or ways, of defining frequency:

- Number of times the word appears in a document (raw count).
- Term frequency adjusted for the length of the document (raw count of occurences divided by number of words in the document).
- Logarithmically scaled frequency (e.g. log(1 + raw count)).
- Boolean frequency (e.g. 1 if the term occurs, or 0 if the term does not occur, in the document).

IDF (inverse document frequency)

Inverse document frequency looks at how common (or uncommon) a word is amongst the corpus. IDF is calculated as follows where ***t*** is the term (word) we are looking to measure the commonness of and ***N*** is the number of documents (d) in the corpus (D).. The denominator is simply the number of documents in which the term, ***t***, appears in.

$$idf(t, D) = log\left(\frac{N}{count(d \in D : t \in d)}\right)$$

Putting it together: TF-IDF

To summarize the key intuition motivating TF-IDF is the importance of a term is inversely related to its frequency across documents.TF gives us information on how often a term appears in a document and IDF gives us information about the relative rarity of a term in the collection of documents. By multiplying these values together we can get our final TF-IDF value.

3eThe higher the TF-IDF score the more important or relevant the term is; as a term gets less relevant, its TF-IDF score will approach 0.

Where to use TF-IDF

As we can see, TF-IDF can be a very handy metric for determining how important a term is in a document. But how is TF-IDF used? There are three main applications for TF-IDF. These are in *machine learning*, *information retrieval*, and *text summarization/keyword extraction.*

Using TF-IDF in machine learning & natural language processing

Machine learning algorithms often use numerical data, so when dealing with textual data or any natural language processing (NLP) task, a sub-field of ML/AI dealing with text, that data first needs to be converted to a vector of numerical data by a process known as vectorization. TF-IDF vectorization involves calculating the TF-IDF score for every word in your corpus relative to that document and then putting that information into a vector. Thus each document in your corpus would have its own vector, and the vector would have a TF-IDF score for every single word in the entire collection of documents. Once you have these vectors you can apply them to various use cases such as seeing if two documents are similar by comparing their TF-IDF vector using cosine similarity.

Using TF-IDF in information retrieval

TF-IDF also has use cases in the field of information retrieval, with one common example being search engines. Since TF-IDF can tell you about the relevant importance of a term based upon a document, a search engine can use TF-IDF to help rank search results based on relevance, with results which are more relevant to the user having higher TF-IDF scores.

Using TF-IDF in text summarization & keyword extraction

Since TF-IDF weights words based on relevance, one can use this technique to determine that the words with the highest relevance are the most important. This can be used to help summarize articles more efficiently or to simply determine keywords (or even tags) for a document.

Vectors & Word Embeddings: TF-IDF vs Word2Vec vs Bag-of-words vs BERT

As discussed above, TF-IDF can be used to vectorize text into a format more agreeable for ML & NLP techniques. However while it is a popular NLP algorithm it is not the only one out there.

Bag of Words

Bag of Words (BoW) simply counts the frequency of words in a document. Thus the vector for a document has the frequency of each word in the corpus for that document. The key difference between bag of words and TF-IDF is that the formetdoes not incorporate any sort of inverse document frequency (IDF) and is only a frequency count (TF).

Word2Vec

Word2Vec is an algorithm that uses shallow 2-layer, not deep, neural networks to ingest a corpus and produce sets of vectors. Some key differences between TF-IDF and word2vec is that TF-IDF is a statistical measure that we can apply to terms in a

document and then use that to form a vector whereas word2vec will produce a vector for a term and then more work may need to be done to convert that set of vectors into a singular vector or other format. Additionally TF-IDF does not take into consideration the context of the words in the corpus whereas word2vec does.

BERT - Bidirectional Encoder Representations from Transformers

BERT is an ML/NLP technique developed by Google that uses a transformer based ML model to convert phrases, words, etc into vectors. Key differences between TF-IDF and BERT are as follows: TF-IDF does not take into account the semantic meaning or context of the words whereas BERT does. Also BERT uses deep neural networks as part of its architecture, meaning that it can be much more computationally expensive than TF-IDF which has no such requirements.

1) Pros of using TF-IDF

The biggest advantages of TF-IDF come from how simple and easy to use it is. It is simple to calculate, it is computationally cheap, and it is a simple starting point for similarity calculations (via TF-IDF vectorization + cosine similarity).

2) Cons of using TF-IDF

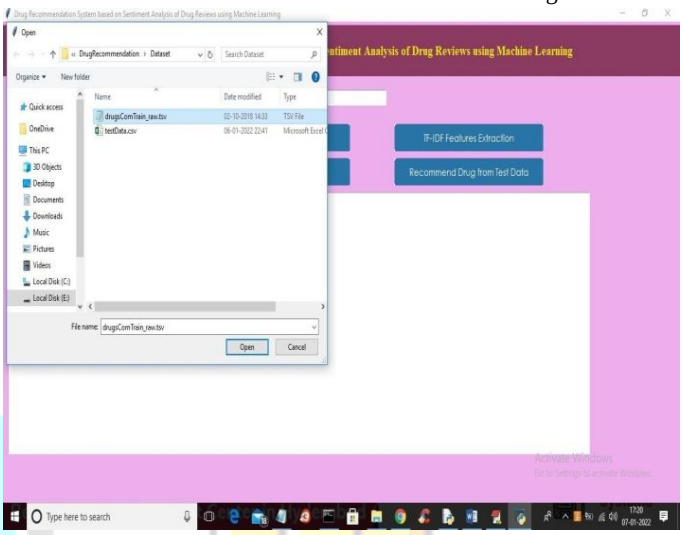
Something to be aware of is that TF-IDF cannot help carry semantic meaning. It considers the importance of the words due to how it weighs them, but it cannot necessarily derive the contexts of the words and understand importance that way.

Also as mentioned above, like BoW, TF-IDF ignores word order and thus compound nouns like “Queen of England” will not be considered as a “single unit”. This also extends to situations like negation with “not pay the bill” vs “pay the bill”, where the order makes a big difference. In both cases using NER tools and underscores, “queen_of_england” or “not_pay” are ways to handle treating the phrase as a single unit.

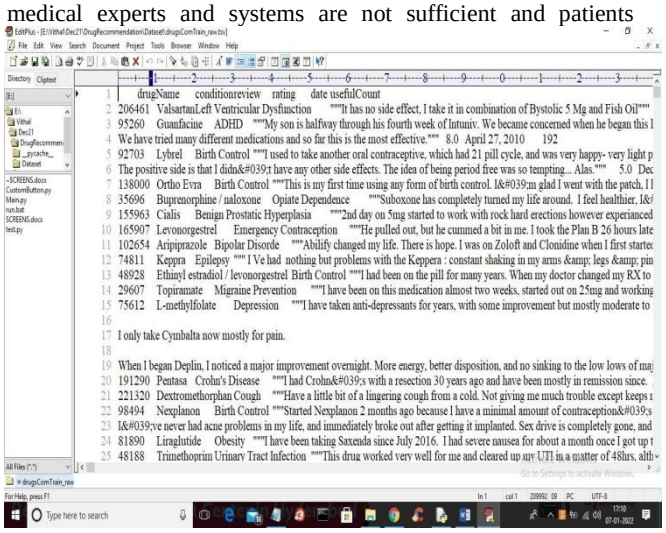
Another disadvantage is that it can suffer from memory-inefficiency since TF-IDF can suffer from the curse of dimensionality. Recall that the length of TF-IDF vectors is equal to the size of the vocabulary. In some classification contexts this may not be an issue but in other contexts like clustering this can be unwieldy as the number of documents increases. Thus looking into some of the above named alternatives (BERT, Word2Vec) may be necessary.

Now-a-days new diseases are attacking human world and corona virus is such diseaseand this diseases require lots of medical systems and medical human experts and dueto growing disease medical experts and systems are not sufficient and patients willtake medicines on their risk which can cause serious death or serious damage topatientbody.To overcome from above problem author of this paper introducing sentiment and machine learning based drug recommendati on system which will accept disease names from patient and then recommend DRUG and simultaneously displayS ENTIMENT rating based on reviews given by old users based on their experience. If predicted rating is high then patientcan trust and took recommended drug.

In propose paper author has used various features extraction algorithms such as TF-IDF(termfrequency–inverse document frequency),BAG of WORDS and WORVEC and this extracted features will be applied on various machine learningalgorithm such as Logistic Regression, Linear SVC, Ridge classifier, Naïve Bayes,Multi layer 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 DRUG REVIEW dataset from UCI machine learning website



and below is th edataset screenshots



In above screen first row represents dataset column names such as drug name, condition, review and rating and remaining rows contains data set 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 read & Preprocess Dataset: using this module we will read all reviews, drug name and ratings from dataset and form a features array.

2) 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 appear in sentence then will be put.All reviews will beconsider as input feature stomach in e learning algorithm and RATINGS and Drug Name willbe consideras class label.

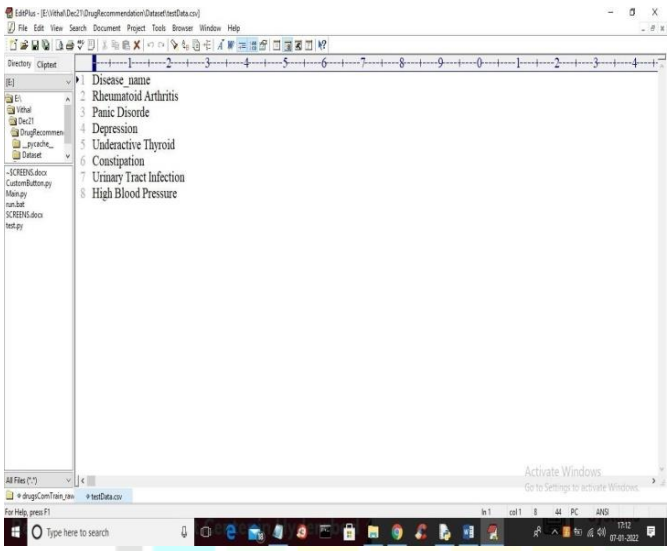
3) Train Machine Learning Algorithms: using this module we willinput TF-IDF features to all machine learning algorithms and thentrained amodel and thismodel will beapplied on test data tocal culate prediction accuracy of the algorithm.

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

Recommend Drug from Test Data:using this module we will upload disease name test data and then ML will predict drug name and ratings.

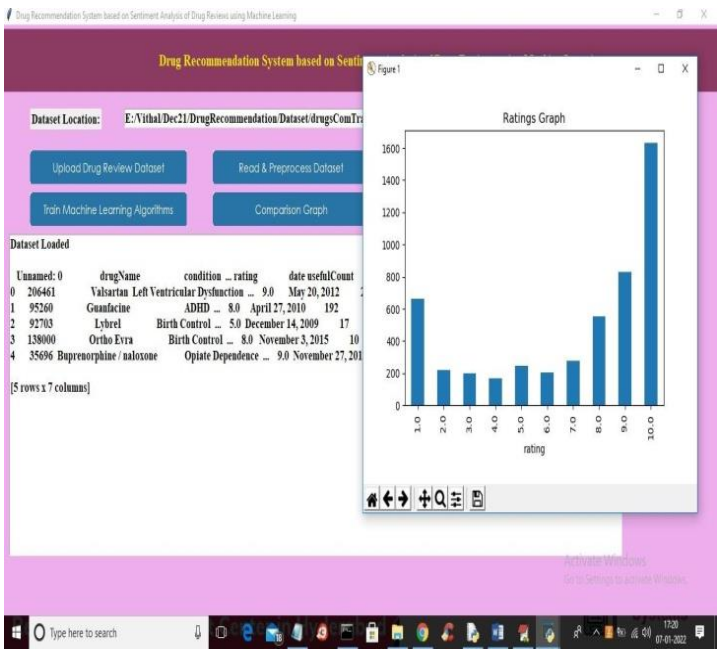
To run project double click on‘run.bat’file to get below screen

In above screen click on‘ Upload Drug Review Dataset’button to upload data set 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 rating and y-axis represents total number of records which got that rating. Now close above graph and then click on ‘ Read & Pre process Dataset ’ button to read all dataset values and then preprocess to remove stop words and special symbols and then form a features array.

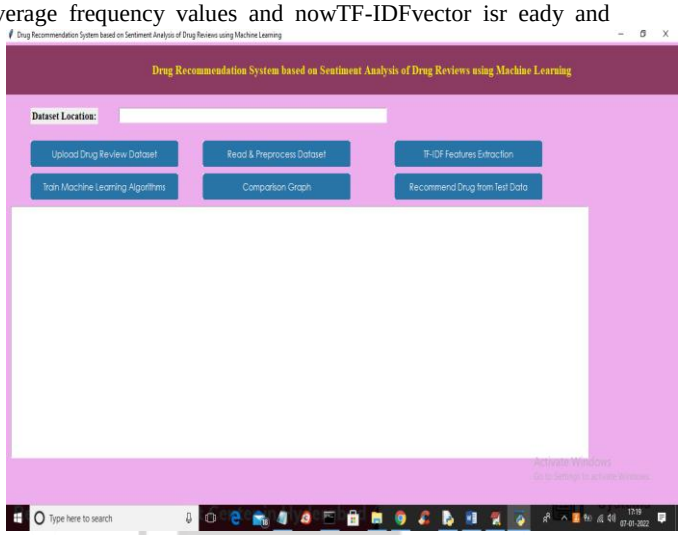


In above screen we can see from all reviews stop words and special symbols are removed and in graph I am displaying

TOP 20 medicinesexist in dataset. In above graph x-axis represents drug name and y-axisrepresents its

count. Now close above graph and then click on ‘TF-IDF Features Extraction’ button to convert all reviews in to average frequency vector

In above screen you can see some columns contains non-zero average frequency values and nowTF-IDFvector isr eady and

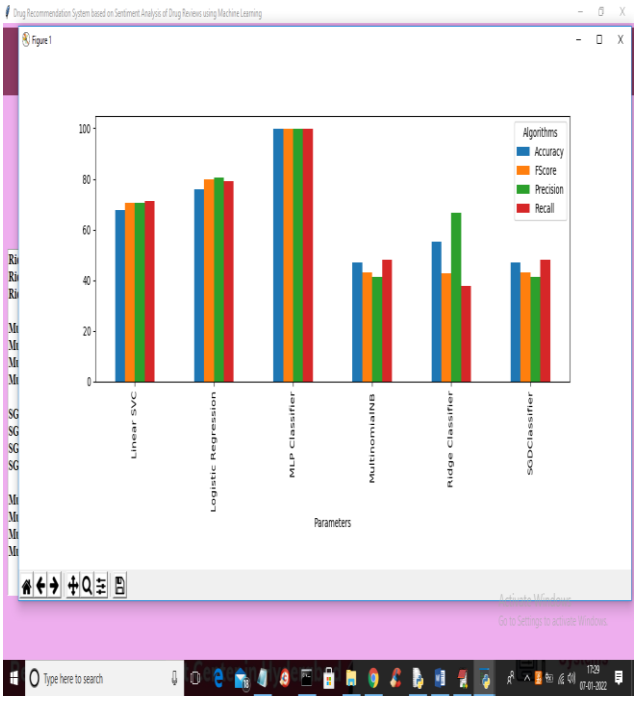


now click on‘Train Machine Learning Algorithm’ button to train all algorithm and get below output

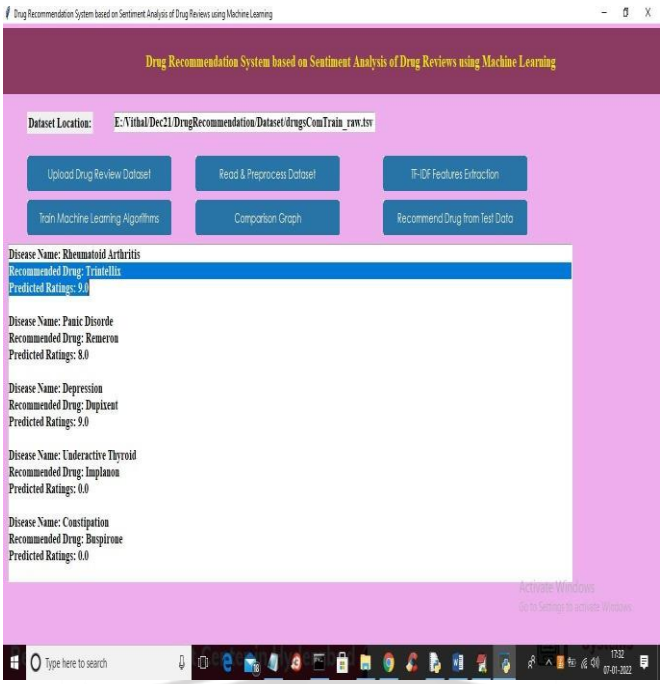
n above screen for each algorithm we calculate accuracy, precision, recall and FSCORE and in all algorithms MLP has got high performance and now clickon‘Comparison Graph’button to get below graph

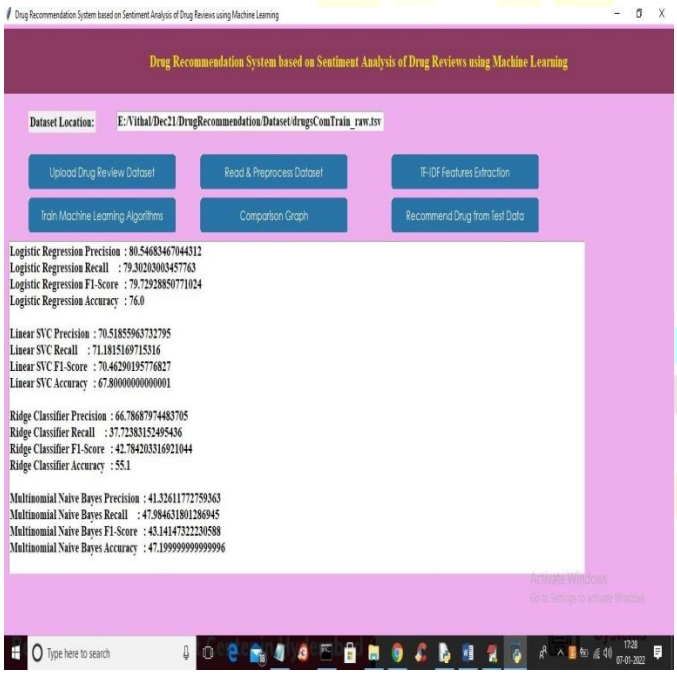
In below graph x-axis represents algorithm name and y-axis represents accuracy, precision recall and FSCORE where each

different colour bar will represents one metric and in above graph we can see MLP got high performance. Now close above graph and then click on ‘Recommend Drug from Test Data’ button to upload test data and to get predicted result as drug name and ratings.



In above screen selecting and uploading 'testData.csv'file and then click on'Open'button to load test data and get below prediction result

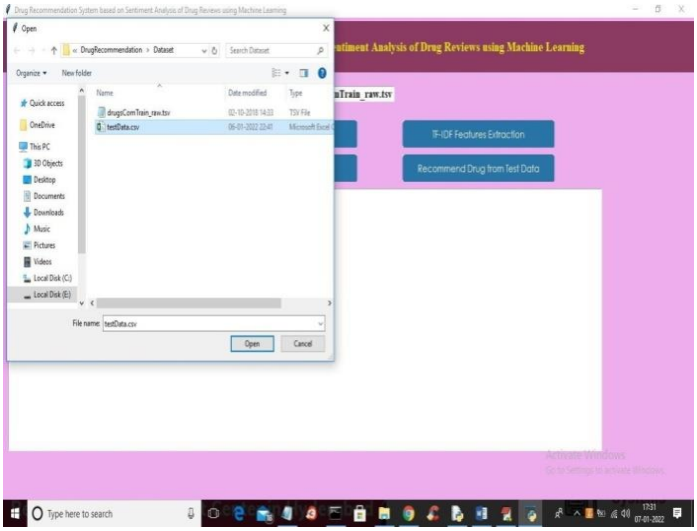




In above screen for each disease name application has predicted recommended drug name and ratings

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,Multinomia lNaive Bayes,Ridge classifier,Stochastic gradient descent, LinearSVC, applied on Bow, TF-IDF, and classifiers such as Decision Tree, Random Forest, Lgbm, and Catboost 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 othermodels with 93% accuracy. On the other hand, the Decision tree classifier on Word2Vecshowed the worst performance by achieving only 78% accuracy. We added best-predictedemotion values from each method, Perceptron on Bow (91%), LinearSVC on TF-IDF (93%),LGBM on Word2Vec (91%), Random Forest on manual features (88%),and multiply themby the normalized useful Count to get the overall score of the drug by condition to build arecommender 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.Medicationerrors:an over view for clinicians.Mayo Clin Proc.2014 Aug;89(8):1116-25.
[3]CHEN,M.R.,&WANG,H.F.(2013).Thereasonandpreventionof hospitalmedicationerrors. PracticalJournalofClinicalMedicine,4.
[4]Drug Review Dataset, https://archive.ics.uci.edu/ml/datasets/Drug%2BReview%2BDataset%2B%2528Drugs.com%2529#
[5]Fox,Susannah,andMaeveDuggan."Healthonline2013.2013."URL:http://pewinternet.org/Reports/2013/Health-online.aspx
[6]Bartlett JG, Dowell SF, Mandell LA, File TM Jr, Musher DM, Fine MJ. Practic eguidelinesf orthemanagemen tof community-acquired pneumoniain adults.Infectious Diseases Society of America. Clin Infect Dis. 2000 Aug;31(2):347-82. doi:10.1086/313954. Epub2000Sep7.PMID:10987697;PMCID:PMC7109923.
[7]Fox, Susannah & Duggan, Maeve. (2012). Health Online 2013. Pew Research Internet Project Report.
[8]T.N.TekadeandM.Emmanuel,"Probabilistic aspect mining approach for interpretation and evaluation of drug reviews,"2016 International ConferenceonSignalProcessing,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-baseddrugrecommendationsdiscovery.JBiomedSemant3,14(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. InProceedings of the 22ndACM SIGKDD International Conference on KnowledgeDiscovery and Data Mining (KDD '16). Association for Computing Machinery, NewYork,NY,USA,1865–1874.DOI:https://doi.org/10.1145/2939672.2939866
[11]V.Goel,A.K.GuptaandN.Kumar,"SentimentAnalysisof MultilingualTwitter Data using Natural Language Processing," 2018 8th International Conferenceon 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 forpatientswith infectious diseases.AMIAAnnu SympProc.2005;2005:1112.
[13]Y. Bao and X. Jiang, "An intelligent medicine recommender system framework,"2016IEEE11thConferenceonIndustrial ElectronicsandApplications(ICIEA),Hefei,2016,pp.1383-1388,doi:10.1109/ICIEA.2016.760380
[14]Zhang, Yin & Zhang, Dafang& Hassan, Mohammad & Alamri, Atif & Peng,Limei. (2014). CADRE: Cloud-Assisted Drug REcommendation Service for OnlinePharmacies. 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 LSTMrecurrentneuralnetworksforhashtagrecommendation,"2016 InternationalJointConference 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-wordsmodel:Astatistical framework.International Journal of Machine Learning and Cybernetics. 43-52.10.1007 /s 13042-010-0001-0.
[17]J. Ramos et al., "Using tf-idf to determine word relevance in document queries,"in Proceedings of the first instructional conference on machinelearning, vol. 242, pp.133–142,Piscataway,NJ,2003.
[18]Yoav Goldberg and Omer Levy. word2vec Explained: deriving Mikolov et al.'snegative-samplingword-embeddingmethod,2014;arXiv:1402.3722.
[19]DanushkaBollegala,TakanoriMaeharaandKenichiKawarabayashi.UnsupervisedCross-DomainWordRepresentationLearning,2015;arXiv:1505.07184.[20]Textblob,https://textblob.readthedocs.io/en/dev/.
[21]van der Maaten, Laurens & Hinton, Geoffrey. (2008). Viualizing data using t-SNE.Journal of Machine Learning Research.9.2579-2605.
[22]N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer. SMOTE:Synthetic Minority Over-sampling Technique, 2011, Journal Of Artificial Intelligence Research,Volume16,pages321-357, 2002;arXiv:1106.1813.DOI:10.1613/jair.953.
[23]Powers,David&Ailab,.(2011).Evaluation:From precision,recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol.2. 2229-398 10.9735/2229-3981 [24] Haibo He, Yang Bai, E. A. Garcia and ShutaoLi,"ADASYN:Adaptive synthetic sampling approach forim balanced learning,"2008IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 1322-1328, doi:10.1109/IJCNN.2008.4633969.[25]Z.Wang,C.Wu,K.Zheng, X.Niu and X.Wang, "SMOTETomek Based Resampling for Personality Recognition,"in IEEE Access,vo