**Department of Information Technology**

**Puducherry Technological University**

**B.TECH (IT) Final Year Mini-Project**



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

***Under the Guidance of***

**Dr. M.EZHILARASAN**

**Professor**

**Department of Information Technology**

**Submitted by: DIVYADARSHINI.G (Reg No:18IT1011)**

**MEIYAZHAGAN@VIKNESH.D (Reg No:18IT1028)**

**NARMITHA.V (Reg No:18IT1032)**

**Problem Definition:**

* Everyday a new study comes up with more drugs,it turns out to be challenging for doctors to choose which treatment or medications to give a patient based on indications and past clinical history.
* Review have become an integral factor. Most proposals are on the E-Commerce field.
* There are number of individuals worried about well-being and finding a diagnosis online**.**

**Motivation:**

* The motivation in this research sentiment analysis of drug reviews is to study and build a recommender system using different types of machine learning classifiers.
* Recommender assist specialists and patients to build knowledge on specific health Conditions.
* Recommender framework proposes an item to the user dependent on their advantage and necessity.
* Framework uses customers and suggest recommend mer surveys to break down their sentiments and suggest recommendation.
* We evaluated them using five different metrics, precision, recall, f1score, accuracy, and AUC score.

**Existing works:**

* Multilingual sentiment analysis was performed using Naïve

Bayes and Recurrent Neural Networks (RNN). Google translator API was used to convert multilingual tweets into the English language. Results show that RNN outperforms Naïve Bayes

**Accuracy :**

**RNN : 95.34%**

**Naïve Bayes : 77.24%**

* Xiaohong Jiang et al. [13] examined three distinct algorithms , Decision Tree Algorithm , Support Vector Machines(SVM) and backpropogation neural networks on treatment data…SVM was picked for the medication proposal module as it performed well in unique boundaries - model exactness, model proficiency, model versatility. Additionally, proposed the mistake check system to ensure analysis, precision and administration quality.
* Mohammad Mehedi Hassan et al. [14] developed a **Cloud assisted drug proposal (CADRE)**. As per patients’ side effects , CADRE can suggest drugs with **top-N related prescriptions**. This proposed framework was initially founded on collaborative filtering techniques in which the medications are initially bunched into clusters as indicated by the functional description data
* Considering the significance of hashtags in sentiment analysis, Jiugang Li et al. [15] constructed a hashtag recommender framework that utilizes the skip-gram model and applied convolutional neural networks (CNN) to learn semantic sentence vectors. These vectors use the features to classify hashtags using LSTM RNN. Results depict that this model beats the conventional models like SVM, Standard RNN.

**Domain Specification:**

**SOFTWARE REQUIREMENTS:**

Operating system : Windows 10

Python IDE : python 3.2.7 , Pycharm

**HARDWARE REQUIREMENTS:**

Ram :4GB and Higher

Processor :i3 processor

Hard Disk : 500GB

**ALGORITHMS:**

* Logistic Regression
* Multinational Naïve Bayes
* Ridge Classifier
* Stochastic gradient Descent
* Linear SVC
* Decision Tree
* Random Forest
* LGBM
* Catboost

**Limitation in the existing System:**

* These vectors use the features to classify hashtags using LSTM RNN. Results depict that this model beats the conventional models like SVM, Standard RNN.
* This exploration depends on the fact that it was undergoing regular AI methods like SVM and collaborative filtering techniques; the semantic features get lost, which has a vital influence in getting a decent expectation.
* Over 40% medicine, specialists make mistakes while prescribing since specialists compose the solution as referenced by their knowledge, which is very restricted.
* Choosing the toplevel medication is significant for patients who need specialists that know wide-based information about microscopic organisms, antibacterial medications, and patients.

**Proposed System:**

**Input:** The dataset used in this research is Drug Review Dataset (Drugs.com) taken from the UCI ML repository. 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.

**Process**

* Data cleaning and visualization removing null values,duplicay and conditions that have no meaning.
* Condition and drug column were joined with review text and Text processing will be done to clean up review text.
* Sentiment Analysis label every single review as positive and negative based on rating
* Machine learning algorithm were used to build a classifier to predict sentiment,which will be measured using metrices
* Overall score of drug of a particular condition will be generated,higher the score ,the better is the drug

**Output:** Drug Recommender based on Sentiment Analysis provide a drug recommendation based on patients history and condition using reviews.

**Design Diagram of the Proposed System:**



**Modules:**

**DATA CLEANING AND VISUALIZATION**

* Data preparation like checking null values,duplicaterows,removing unnecessary values and text from rows.
* Condition and drug colum joined with review text because condition medication also have predictive power.text processing will be done and review transferred to lowercase.
* Sentiment analysis label every single review as positive and negative based on its user ratings.
* Data sets will be cleaned and processed to make proper set up of the data required to build classifiers for sentiment Analysis.

**FEATURE EXTRACTION**

Feature extraction will be done by using techniques like BOW,TF-IDF,Work2Vec and Manual Feature.

**BOW:** Bag of words[16]is an algorithm used in NLP for counting number of tokens in review.It consider all the terms in the corpus which in turn build a large matrix that is computationally expensive.

**TF-IDF:**TF-IDF[17] weighing stratergy offers words weight not count.Gives low importance to words that appears often.TF can be called likelihood of locating a word in a document.

tf(t,d) = log(1 + freq(t,d))

IDF is opposite of the number of times a specific term in corpus

Idf(t,d) = log (N/count(dϵD : tϵd)

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

tfidf(t,d,D) = tf(t,d).idf(t,D)

**Worrk2Vec**[18]**:**Used toProduce word embedding which reproduced from deep learning models.Take the semantic meaning of words and arrange vvector of words in vector space.Words with similar sense found in dataset are found close in one another in vector space.

**Manual Feature**:Helps to increase the accuracy of model.Textblob toolkit[20] was used to extract the cleaned and uncleaned riews polarity and added as feature with a a 8 features generated from each of the reviews.

**TRAIN TEST SPLIT**

* There are 4 types of Datasets using Bow,TF-IDF,word2Vec and manual features
* And they were split into 75% Training and 25% Testing randomly
* SMOTE (Synthetic minority Over-sampling technique).
* After test and train split, only training data will undergo **SMOTE** in order to prevent the class imbalance problem
* Smote is an over-sampling technique that synthesized new data from existing data

**CLASSIFIERS**

* A Machine learning algorithm that automatically orders or categorizes data into one or more of a set of classes
* Classification algorithms were used to build a classifier to predict the sentiments
* Logistic Regression,Multinomial Naïve Bayes,Perceptron,Ridge classifier experimented with Bow ,TF-IDF model since ,they are very sprase matrix
* Appling tree-based classifiers like Decision tree,RandomForest,LGBM on Word2vec and manual features model were time consuming
* We selected those machine learning algorithms only that reduces the training and give faster predictions

**METRICS**

* The predicted sentiment were measured using 5 metrics:
* Precision
* Recall
* Accuracy
* F1 Score
* AUC Score[17**]**

**Precision**

Precision is the ratio between the **True Positives** and **all the Positives**

**Precision = True Positive (TP) /True Positive (TP) + False Positive(FP)**

**Recall**

The recall is the measure of our model correctly identifying True Positives

Recall = TP / TP + FN

**Accuracy**

Accuracy is the **number of correct predictions made by the model** over all kinds of predictions made

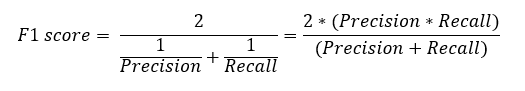
Accuracy = TP + TN / TP + FP + FN + TN

In the Numerator, are our **correct predictions** (True positives and True Negatives) and in the denominator, are the kind of **all predictions made by the algorithm**

**F1 Score**

It is the harmonic mean of precision and recall

It takes both FP and FN into account , therefore It performs well on imbalanced dataset.

****

**AUC Curve**

The **Receiver Operator Characeristic curve (ROC)** is an evaluation meric for binary classification problems.

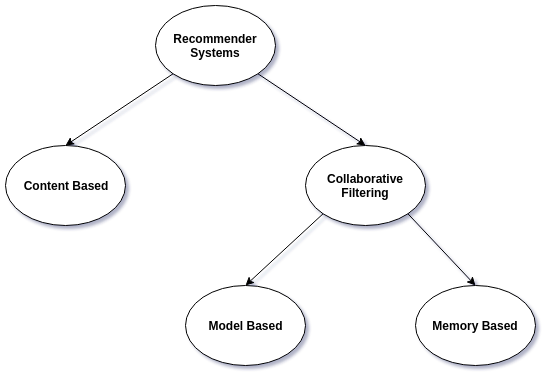
It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’

The **AUC (Area under the Curve)** is the measure of the ability of a classifier to distinguish between the classes and is used as a summary of he ROC curve

**RECOMMENDER SYSTEMS**

* The recommender system deals with a large volume of information present by filtering the most important information based on the data provided by a user and other factors that take care of the user’s preference and interest.
* It finds out the match between user and item and imputes the similarities between users and items for recommendation.
* The high level idea behind an item recommendation system is to rank a large catalog of items and show only top most relevant items to the user
* Machine learning algorithms in recommender systems are typically classified into two categories :
* content based
* collaborative filtering methods
* Although modern recommenders combine both approaches , Content based methods are based on similarity of item attributes and collaborative methods calculate similarity from interactions.

**A tree of the different types of recommender system**

****

**COLLABORATIVE FILTERING**

* Collaborative filtering methods for recommender systems are methods that are solely based on the past interactions between users and the target items. Thus, the input to a collaborative filtering system will be all historical data of user interactions with target items. This data is typically stored in a matrix where the rows are the users, and the columns are the items.
* The core idea behind such systems is that….. the historical data of the users should be enough to make a prediction. i.e we don’t need anything more than that historical data, no extra push from the user, no presently trending information, etc

**DRUG RECOMMENDER SYSTEMS**

* After assessing the metrics, predicted results were picked and joined together to produce the combined prediction
* Merged results \* normalised useful count = overall score of drug of a particular condition
* 
* Higher the score , better the drug.
* In this work, each review was classified as positive or negative, depending on the user’s star rating.
* Ratings above five are classified as positive, while negative ratings are from one to five-star ratings**.**

**CONCLUSION**

* 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%), LinearSVC on TF-IDF (93%), LGBM on Word2Vec (91%), Random Forest on manual features (88%), and multiply them by the normalized usefulCount to get the overall score of the drug by condition to build a recommender system.
* optimization of algorithms to improve the performance of the recommender system.

**TIMELINE CHART**

**Gantt chart (Aug 2021 - Jan 2022)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MINI-PROJECT** | |  | | | | |  |
| Aug | Sep | Oct | Nov | Dec | Jan |
| **ACTIVITIES** | Identify the Problem |  |  |  |  |  |  |
| Study of Existing work |  |  |  |  |  |  |
| Define the Problem |  |  |  |  |  |  |
| Develop a model |  |  |  |  |  |  |
| Propose Methodology |  |  |  |  |  |  |
| Describe the Modules |  |  |  |  |  |  |
| Collect the Data |  |  |  |  |  |  |
| Implement the Modules |  |  |  |  |  |  |
| Code the Modules |  |  |  |  |  |  |
| Test the Proposed System |  |  |  |  |  |  |
| Prepare the Report |  |  |  |  |  |  |

**Conclusion**

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%), LinearSVC on TF-IDF (93%), LGBM on Word2Vec (91%), Random Forest on manual features (88%), and multiply them by the normalized usefulCount to get the overall score of the drug by condition to build a recommender system.

optimization of algorithms to improve the performance of the recommender system.

**BASE PAPERS:**

[1]Satvik Garg, “Drug Recommendation System based on Sentiment Analysis of Drugs Reviwes using Machine Learning” Jaypee University of InformationTechnology,IEEE|DOI:10.1109/Confluence51648.2021.9377188

[2] J.Shobana , M.Murali “ An efficient sentiment analysis methodology based on LSTM “

SRM Institute of Science and Technology , Chennai , India , Springer | Complex & Intelligent

Systems(2021) 7 :2485-2501

[3]Thi Ngoc Trang Tran, Alexander Felfernig,Chistoph Trattner, Andreas Holzinger “Recommender systems in the healthcare domain: state of the art and research issues”

**REFERENCES:**

[1] 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.

[2]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.

[3]Leilei Sun, Chuanren Liu, Chonghui Guo, Hui Xiong, and YanmingXie. 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.

[4] Zhang, Yin & Zhang, Dafang& Hassan, Mohammad & Alamri, Atif &Peng, Limei. (2014). CADRE: Cloud-Assisted Drug RecommendationService for Online Pharmacies. Mobile Networks and Applications. 20. 348-355. 10.1007/s11036-014-0537-4.

[5] J. Li, H. Xu, X. He, J. Deng and X. Sun, ”Tweetmodeling 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.

[6] van der Maaten, Laurens & Hinton, Geoffrey. (2008). Viualizing data using t-SNE. Journal of Machine Learning Research. 9. 2579-2605.

[7]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, Volume 16, pages 321-357, 2002

[8] Powers, David &Ailab,. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2. 2229-3981. 10.9735/2229-3981

[9] Haibo He, Yang Bai, E. A. Garcia and Shutao Li, ”ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

[10] Z. Wang, C. Wu, K. Zheng, X. Niu and X. Wang, ”SMOTETomek- Based Resampling for Personality Recognition,” in IEEE Access, vol. 7, pp. 129678-129689, 2019, doi: 10.1109/ACCESS.2019.2940061

[11] 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.

[12] 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.

[13] Yoav Goldberg and Omer Levy. word2vec Explained: deriving Mikolov et al.’s negative-sampling word-embedding method, 2014; arXiv:1402.3722.

[14] Danushka Bollegala, Takanori Maehara and Kenichi Kawarabayashi. Unsupervised Cross-Domain Word Representation Learning, 2015; arXiv:1505.07184.

[15] Textblob, https://textblob.readthedocs.io/en/dev/.

[16] 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, Volume 16, pages 321-357, 2002; arXiv:1106.1813. DOI: 10.1613/jair.953.

[17] Powers, David &Ailab,. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2. 2229-3981. 10.9735/2229-3981