Medicine Recommendations Using Sentimental Analysis of Drugs Reviews

Rutik Rajeshkumar Kothwala
CS668, Analytics Capstone
PACE University
New York, USA
rk29015n@pace.edu

***Abstract*—Sentimental Analysis has been acquiring a crucial role in both commercial and research application as one of the most emerging subfields of natural language processing. Most of the work on sentimental analysis is on Movie reviews or for an Ecommerce platform. The following study is an attempt to investigates use of machine learning and natural language processing for sentimental analysis on drugs review to create a drug recommendation system framework which recommend most effective drugs for the given condition by analyzing drugs descriptions, patients' condition, and reviews which will be not only useful for patients but also for pharmacy companies and clinicians to improve consumer safety by assisting in reduction of medical errors. For this work data set on drugs reviews is taken from UCI machine learning repository our study includes text-preprocessing task such as stop word removal, tokenization, lemmatization, feature engineering techniques TF-IDF, bag-of-words, word embedding, sentiment classification baseline machine learning models Naïve Bayes, Random Forest, XGBOOST and deep learning models N-gram, LSTM. Evaluation metrics to be consider are Accuracy, F1-Socre and Confusion Matrix, goal is to get data insights and drug recommendation with accuracy of 80%.**

Keywords—Drugs Reviews, Sentimental Analysis, Feature Extractor, Machine Learning, Deep Learning and Recommendation System.

I. Introduction

These days, everything is accessible on the internet. Customers go online for product reviews and comments before making any kind of purchase. People may eventually become weary of hearing whether a product is better based on remarks.

A recommendation system is a class of machine learning that uses data to help predict, narrow down and find what people are looking for among an exponentially growing number of options from a vast amount of complicated knowledge with the development of social media anyone may now readily express their opinions. Sentimental Analysis is a Natural Language Processing (NLP) technique that categorizes opinions from pieces of text to determine a sentiment score (positive, negative, or neutral) to better understands a person's reaction and attitudes, towards several entities.

User-generated reviews, especially those about medications, are abundant due to quick expansion of social media and websites that solicit feedback from users, they are helpful to regulatory agencies, pharmaceutical companies, and healthcare professionals. Drug reviews are necessary to detect and treat drug-related diseases, which will enhance medication use and health outcomes. Also text mining on drug reviews will be useful to face a major challenge of medication errors. Medication errors are one of the most serious medical errors that could threaten patients' life. More than 42% of these errors are caused by doctors who have limited experiences/knowledge about drugs and diseases. But a major challenge is the enormous volume of evaluations and specialized medical knowledge needed. It is not feasible to go through each of these comments by hand.

Machine learning techniques have proven to be effective in analyzing large volumes of data and extracting meaningful insights. Therefore, sentimental analysis and machine learning models can be used on drugs reviews to prepare desired recommendation framework.

II. Literature Review

A. Current Medical Landscape

With the growing demand for medical services, there is a huge burden on healthcare providers as there is a shortage of medical practitioners. Even though the World Health Organization (WHO) suggests a ratio of 1:1000 doctor population ratio, the workload on medical professionals has been never-ending.

It is a complicated procedure for a medical professional to administer medication therapy to a patient. From the prescription process until the final delivery of the medication to the patient, mistakes might happen at any stage. Medication errors are frequently caused by misdiagnoses, prescription errors, dose calculations gone wrong, improper drug distribution procedures, issues with drugs and drug devices, improper drug administration, poor communication, and a lack of patient education.

Incorrect prescriptions are one of the main reasons for therapeutic pharmaceutical mishaps. Between 1995 and 2000, there were 198,000 fewer patient deaths because of medication errors than there were in 2000 (218,000). The US economy suffers losses from these follies totalling over $177 billion annually.

With the advances and advantages of Artificial Intelligence (AI) technologies, there is an opportunity to tackle these problems with the help of Machine Larning and Deep Learning

B. Previous Study and Learning from it

Sentiment Analysis using product review data Xing Fang & Justin Zhan in it experiments for both sentence level categorization and review level categorization are performed with promising outcomes from it ,came to learn which types of classification model and evaluation metrics can be used for the research they use Naïve Bayes , Random Forest, XgBoost and Support Vector Machines (SVM) and F1 score as most focus evaluating metrics obtaining best performance for SVM.

Sentimental Analysis in social media and its applications: Systematic Literature Review by Zulfadzli Drus, Haliyana Khalid. This paper focuses to provide a better understanding of the application of sentiment analysis in social media platform by examining related literature learn There are 2 main methods of sentiment analysis have been identified which is a machine learning approach and lexicon-based approach. Machine learning approach utilized algorithms to extract and detect sentiment from a data while lexicon-based approach works by counting the positive and negative words that related to the data. Research is going on developing a new effective and accurate model in sentiment analysis.

Comparing deep learning architectures for sentiment analysis on drug reviews by Cristóbal Colón-Ruiz and Isabel Segura-Bedmar they present a benchmark comparison of various deep learning architectures such as Convolutional Neural Networks (CNN) and Long short-term memory (LSTM) recurrent neural networks and propose several combinations of these models and also study the effect of different pre-trained word embedding models. Learn that pretrained models like BERT gives the best result but with a very high training time. On the other hand, LSTM achieves acceptable results while requiring less training time, standard evaluation metrics where taken for text classification tasks: precision, recall, accuracy and F1.

Drugs reviews sentiment analysis using weakly supervised model by Zhang Min. Firstly, this paper proposes a weakly supervised mechanism (WSM) that applies the weakly labelled data to pre-train the parameters of the model then use labelled data to fine-tune initialed parameters, which reduces the effect of noise data on the consequences. Secondly, present a novel architecture apply the WSM and combines the strength of Convolutional Neural Network (CNN) and Bi-directional Long Short-term Memory (Bi-LSTM) named as WSM-CNN-LSTM to complete the task for drug reviews sentiment classification and learned that that the weakly supervised mechanism only required a small amount of labelled data to acquire optimal performance, which reduces the manual data-labelling requirements because lower the number of labelled data required lower the manual labor needed.

Sentimental Classification of Drug Reviews Using Machine Learning Techniques by Mohammad Al-Ameen A. Hameed; Khalid Shaker; Haitham A. Khalaf. in this paper, new approaches have been developed that are based on patient reviews to predict sentiment to improve data analysis. They use Term Frequency-Inverse Document Frequency (TF-IDF) and CountVectorizer. After the raw reviews have been preprocessed, each tokenized review is acted as a length matrix equal to the number of the distinct phrase in the returned corpus. The tf-idf and countvectorizer technique is used to determine the value of each word in the related document. Their experimental findings show that the Random Forest Classifier (RFC) beats all results of other existing models from the literature in terms of Precision, Recall, F1-Score, and Accuracy of 93 % accuracy.

Getting Started with Sentiment Analysis using Python by Federico Pascual on Huggingface is an amazing source for anyone who want to learn sentimental analysis with python in it what is sentimental analysis?, how to use pretrained sentimental analysis models with python, how to build your sentimental analysis model, fine tuning everything is explained in detail with code I am referring this site for the given project, moreover in it various example with codes are also given like analyzing IMDB movie reviews, analyzing tweets etc.

III. Research Methodology

The Proposed architecture for our research is as follows.

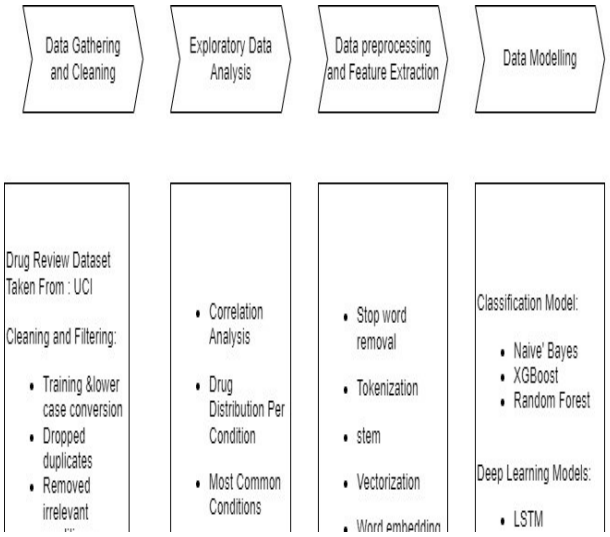

Fig. 1. Diagram of Proposed Framework

A. Data Gatherings

This step involves the gathering of suitable data from different sources, collected data must be diverse and representable. For our research, drug review dataset is selected from UCI Machine Learning Repository.

The UCI ML Drug Review dataset provides patient reviews on specific drugs along with related conditions and a 10-star patient rating system reflecting overall patient satisfaction. The data was obtained by crawling online pharmaceutical review sites.

Dataset Characteristics	Subject Area	Associated Task	Feature Type	Instances	Missing Values
Multivariate, Text	Health and Medicine	Sentimental Analysis, Predictive Modelling	Integer	215063	No

Fig. 2. Dataset Charaacteristics

B. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an analysis approach that identifies general patterns in the data. These patterns include outliers and features of the data that might be unexpected. EDA is an important first step in any data science project.

In our research EDA help to determine different types of drugs, drugs distribution per condition, most common condition and to perform correlation analysis.

C. Data Pre-Processing and Feature Extraction

There are 2 main methods of sentiment analysis that have been identified which is a machine learning approach and lexicon-based approach for our project we will use machine

learning approach. The first step in it is to preprocess the text, this process will make the unstructured data containing noise in such a form that can be used for classification. Preprocessing involves tasks such as tokenization, stop word removal, lower case conversion, stemming and removing numbers.

We will preprocess our data by removing the missing values, removing stopping words, as well as perform text normalization and lemmatization and spelling corrections to reduce the dimensionality of the data. However, care must be taken during the noise removal and text normalization processes, as they can result in the loss of a small number of rows from the dataset, potentially reducing the accuracy.

Next stage is to features extraction. There are different types of text features such as count vectors, bag of words, TF-IDF, word embeddings. We will use two feature extractor techniques TF-IDF and CountVectorizer.

Words within a text document are transformed into important numbers by a text vectorization process. There are many different text vectorizations scoring schemes, with TF-IDF and CountVectorizer being one of the most common.

As now data text data is converted to number it is ready to train the model.

D. Data Modeling

After Data-Preprocessing and Feature Extraction the next step is Data Modeling. There are two approaches for model development, supervised learning, and unsupervised learning. The biggest difference between supervised and unsupervised machine learning is the type of data used. Supervised learning uses labeled training data, and unsupervised learning does not. More simply, supervised learning models have a baseline understanding of what the correct output values should be. In contrast, unsupervised learning algorithms work independently to learn the data's inherent structure without any specific guidance or instruction.

In our work we will use supervised learning algorithms with two different techniques baseline machine learning classification and deep learning.

Supervised learning using ML	Supervised learning using DL
Naïve Bayes	LSTM
Random Forest	
XGBOOST	

Fig. 3. Proposed Model

Supervised learning using Machine Learning: We will use Random Forest, Naïve Bayes and XGBOOST. The reviews in text form transformed into numbers will be used to fed in this classifier model. The main limitation of this classifier model is the low accuracy of these algorithms in NLP tasks but still it is used in various sentimental analysis research and studies due to its less complex interpretation and less training time of these algorithms. To train our ML model, we will split the dataset into training and testing with 80% of the data being used for training and 20% for testing, training test consist of labeled drug reviews and testing set consist of unlabeled reviews.

Overall, the combination of non-linearity handling, robustness to overfitting and scalability makes these 3 models a popular choice for sentimental analysis tasks, including for drugs reviews.

Supervised learning using DL: Neural networks offer advantages when use on a large-scale data like ours 215063 drugs reviews, it has good adaptability to handle complex data, moreover it gives more freedom to adjust hyperparameter and result in overall improving accuracy the basic classification model. The main disadvantage is that it takes longer to compute than other algorithms. There are numerous RNN-based model variations that work well with sequential input data, including audio, music, text, name entity recognition, etc. For our project we will use LSTM as in most of the research works LSTM stands to be the best RNN algorithm, LSTM is a particular kind of RNN that can learn long-term dependencies. We will split our dataset into 80% training, 5% validation, and 20% testing set to train our LSTM deep learning model.

E. Evaluation & Results

Here we will compared three baseline classification model Random forest ,XgBoost and Naïve Bayes and deep learning model LSTM on evaluation metrics Accuracy, F1 score and confusion matrix with the prime focus on accuracy aim to achieve at least 80% of accuracy on our best performing model. In other research works F1 score is selected as main metrics but in our works we have prime focus on accuracy as our model is not just a classification model but it will used to recommend medicine so for us accuracy is important.

The resulting model can then be used to recommend the most effective drugs based on the given condition of the patients.

CONCLUSION

In conclusion, this work investigates the use of sentiment analysis in drug reviews to create a reliable drug recommendation system. Our goal is to provide consumers, pharmacy firms, and clinicians with useful medication recommendations for illnesses by utilizing machine learning and natural language processing techniques. We use many categorization models, engineer features, and preprocess text data. We want to have an accuracy rate of 80%. This study emphasizes the potential of sentiment analysis in the medical field and the need for ongoing improvement to maximize patient safety and care.

REFERENCES

[1] Sentimental Analysis in social media and its applications: Systematic Literature Review by Zulfadzli Drus, Haliyana Khalid.

[2] Sentiment Analysis using product review data Xing Fang & Justin Zhan Journal of Big Data, Article number 5.

[3] Comparing deep learning architectures for sentiment analysis on drug reviews by Cristóbal Colón-Ruiz and Isabel Segura-Bedmar from Journal of Biomedical Informatics.

[4] Drugs reviews sentiment analysis using weakly supervised model by Zhang Min, publisher IEEE.

[5] DLRS: Deep Learning-Based Recommender System for Smart Healthcare Ecosystem.by, Gagangeet Singh Aujla; Anish Jindal; Rajat Chaudhary; Neeraj Kumar; Sahil Vashist; Neeraj Sharma; Mohammad S. Obaidat, publisher IEEE.

[6] Deep Learning based Chatbot Architecture for Medical Diagnosis and Treatment Recommendation.by Girish Rajani, Khusi Ruparel, publisher IEEE.

[7] National Library of Medicine-National Healthcare Quality and Disparities Report.

[8] Getting Started with Sentiment Analysis using Python by Federico Pascual on Huggingface.

[9] Sentimental Classification of Drug Reviews Using Machine Learning Techniques by Mohammad Al-Ameen A. Hameed; Khalid Shaker; Haitham A. Khalaf.