

Recommendation System for Drugs and Sentiment Analysis on Drug Reviews

Mounica Subramani, Sai Divya Sangeetha Bhagavatula, Sushma Suresh Kalkunte

Problem Description

Opinion forums and review sites contain an abundance of data regarding user experiences and preferences over a variety of products and domains. A wealth of useful insights can be generated using this data. In this project, we propose to perform sentiment analysis in the healthcare sector, by examining the online reviews of various drugs provided by patients. The project also intends to build a recommendation system to suggest suitable medicines for specific ailments using the reviews of patients.

Dataset

Dataset is derived from Kaggle: <https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018>

The data contains patient reviews on specific drugs along with related conditions and a 10-star patient rating reflecting overall patient satisfaction. Some useful attributes of the data include the name of the drug and condition experienced, reviews of patients, ratings and the number of people who found the reviews useful. The total number of records in the dataset are 215,063. The dataset has been split as 75% for training and 25% for testing. The dataset contains seven features: uniqueID, drugName, condition, review, rating, date (Date of review entry), usefulCount (Number of users who found review useful)

Approach and Methodology

We plan to start by building a recommendation system to suggest suitable medicine for the particular ailment using the previous patients' reviews. This can be achieved by performing text analysis on the patients' reviews and ratings of the drug obtained from historical data, to recommend the most suitable medicine for a new patient with similar symptoms.

We will then proceed to perform sentiment analysis on the reviews, using the below machine learning algorithms:

- Logistic Regression
- Neural Networks
- Light Gradient Boosting Machine(LGBM)

The language that we intend to use for coding is python. Required packages would be - Numpy, pandas, Scikit, matplotlib, tensorflow and keras.

We will conclude the analysis by comparing and evaluating the above mentioned models based on the following metrics-

- F1 score
- Precision
- Recall
- Accuracy