W4995-AML Project Deliverable #1 - Project Proposal

Team 25: Huaizhi Ge, Shashwat Singh, Yosha Singh Tomar, Alex Kita

Background and context to the problem statement:

With the advent of new drugs everyday, it is important to take into account the level of user satisfaction and the effectiveness of the drug. There are also popular websites such as WebMD where users provide feedback regarding the drug usage and effectiveness. We can exploit this data using NLP techniques to determine whether the drug would prove to be beneficial or not. Further, this data along with the historical data which contains personal attributes of various patients can be taken into account by the doctors to recommend a drug to the patient using ML techniques. Such a process would help the doctors as well as the patients in getting the right treatment for their medical condition.

The problem statement is as follows:

1. Using patient reviews, predict the sentiment rating of the text. This would involve rating the reviews on the scale of 1-10 to determine whether the drug proved useful for the patient or not.
2. From historical usage of drugs and feedback of patients on effectiveness, satisfaction and ease of use, recommend drugs to new patients by utilizing information about their existing conditions and other personal attributes, such as age, sex, etc.

Identification and description of the data set(s) you are planning on using along with their source.

- Dataset 1: UCI Drug Review Dataset
  https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29
- Dataset 2: WebMD Drug Reviews Dataset
  https://www.kaggle.com/rohanharode07/webmd-drug-reviews-dataset

Proposed ML techniques you are proposing on applying to solve the problem:

- Task 1:
  - Using the first dataset, train an NLP model (using BERT/ Word2Vec embeddings) to perform sentiment analysis and predict the user rating.
  - Employ this trained model to generate the user satisfaction level in the reviews from the second dataset.
- Task 2:
  - On the second dataset, utilize features such as age, gender, condition and ratings in addition to the predicted user satisfaction level from the NLP model to predict appropriate drugs. This can be implemented using any one or both of the following techniques:
    - Multiclass Classification (Supervised Learning: Logistic Regression/SVMs/Decision Trees & Random Forests/Multi Level Perceptrons)
    - Recommendation Engine
- Task 3 (if time permits, ambitious goal):
  - Enhance the model from the second task using unsupervised clustering techniques. Cluster the independent variables and include it as an additional feature for the supervised model.