**IST 736 – TEXT MINING**

**PROJECT PROPOSAL**



**Drug Review Analysis**

Submitted by:

Akshay Bhala, Sharvil Turbadkar, Yeswanth Reddy Velapalem

**Problem Statement**

Online review sites contain a lot of information regarding user preferences and experiences over multiple product domains which can be used to obtain valuable insights. In this project we examine online user reviews within the pharmaceutical field which contain information related to multiple aspects such as effectiveness of drugs and side effects, which make analysis very interesting but also challenging. However, analyzing sentiments concerning the various aspects of drug reviews can provide valuable insights, help with decision making and improve monitoring patient health. In this preliminary work we intend to perform multiple tasks such as supervised classification and sentiment analysis over drug reviews with data obtained by crawling online review sites. We plan to perform sentiment analysis to predict the sentiments concerning overall rating of user reviews on specific drugs. We further plan to investigate the data by training classification models among top 10 conditions and data sources. In this work we show that text mining approaches can be used to exploit similarities across the reviews of drugs related on different conditions.

**Problem Modeling**

The project can be modeled in two ways 1.Supervised Classification and 2. Sentiment Classification.

**Supervised Classification:**

The project proposes to build a classification model and predict the condition (target variable) of the patients based on the reviews given. Since there are more than 850 unique conditions which makes the problem complex, we are planning to take a subset of the data which comprises of top 10 conditions namely 'ADHD', 'Acne', 'Anxiety', 'Bipolar Disorder', 'Birth Control', 'Depression', 'Insomnia', 'Obesity', 'Pain', 'Weight Loss'. These 10 conditions will act as the categories. The data subset which is extracted shows imbalance in the classes as Birth control constitutes almost 36% of the data followed by Depression which amounts to 12%. All the other classes comprise, 4-8% each, of the data.

**Sentiment Classification:**

We also intend to perform Sentiment Analysis on the reviews by labelling the ratings (target variable) 8-10 as positive, 5-7 as neutral and 1-4 as negative. Our main goal is to check the effectiveness of using Sentiment Analysis which could detect the sentiment of the review and hence be in agreement with the rating classification. There are multiple reviews for the drugs that belong to a similar condition and we decided to investigate how the reviews for different conditions use different words impact the ratings of the drugs.

**Methodology**

For exploratory data analysis, we are going to start with visualizing the conditions and drugs since drugs are closely related to conditions. Word clouds are will be used to find the most common conditions and most prescribed drugs. Then we will classify ratings 1-4 as negative, 5-7 as neutral, 8-10 as positive and use unigrams and n-grams to plot most frequent words to check which corpus best classifies emotions. Pre-processing will include dealing with missing values, dropping unnecessary features and observations, cleaning the reviews using regex. For the classification, we plan on using Multinomial Naïve Bayes , Bernoulli Naïve Bayes and SVM for multi-class classification of drug review analysis with various combinations of vectorization options. We will use the holdout test method and cross validation methods and evaluate the performances. Depending upon the classification algorithm we can use different ways to count words:

1] Boolean (Presence and Absence of a word in a document)

2] Term Frequency (Counts the frequency of a word occurring in a document)

3] Normalized Term Frequency (Term frequency of words divided by the total word count of that document)

4]TF-IDF- Penalizes a word if it occurs in most of the documents.

In feature extraction words will be stemmed or lemmatized — words in third person are changed to first person and verbs in past and future tenses are changed into present. Words will be stemmed to their root form.

**Data Description**

Dataset consists of reviews of different drugs with their condition and a 10-start patient rating which reflects the overall efficiency of drug. This data is collected by crawling online pharmaceutical reviews sites and download from UCI repository [1]. The dataset also contains two additional variables named useful count and date which denotes the count of people who founds the review useful and the date on which the review was submitted. Dataset contains more than 200,000 reviews.

[1]<https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29>

**Evaluation methods**

For evaluation of multi-class classification models, we generally use the accuracy, recall, precision and f1 score. If the data is balanced, accuracy can be used as a reliable evaluation metric. Since we are dealing with an imbalanced dataset, we should consider the macro recall, macro precision and macro f1 score for evaluation of the model performance. We will also use hold-out method and cross validation techniques to compare the performances of the models. The model accuracy should be more than the majority vote baseline. Also, error analysis and model interpretation of the most informative features that models have learned might be helpful for improving model’s performance.

**Expected Challenges**

Selecting a better classification model is going to be a challenge since different algorithms gives different results and the mode of classification of each model is unique. Also, there are multiple vectorization options in addition to the models. Since it is a multiclass classification some of the classes are predicted better by one model and some by another model. Many patients have different sensitivities to different drugs as their body composition is different. Engineering a generalized model that has the most positive effects is the drug of our choice as our aim is to maximize positive effects.