Predicting Drug Effects using text analytics Technical Report

Team Members:

Sakshi Sharma, Ganagadhar Vutla, Madhusudhan Vaddi , Jyoti Thakral, Karthik Niyogi

To do text and sentiment analysis on the drug data, collected by web crawling from the website and analyzing it on the basis of various descriptive parameters. The data was collected on the basis of one specific drug and its related customer reviews. Initially data was plotted to understand the trends visually using tableau. Later the reviews of the drugs were grouped based on their impact on children and overall efficiency. We found a correlation among variables like age groups, side effects of a particular drug, effectiveness and satisfaction. For this process we used Excel, R, SAS and Tableau were used.

Sources:

1. Stop Words:
   http://xpo6.com/download-stop-word-list/

2. List of keywords for sentiment analysis:
   https://github.com/williamgunn/SciSentiment/blob/master/positive-words.txt

3. Link to the website:
   http://www.webmd.com/

Phases of our analysis:

1. **User Interface**: We created a web interface for where we allow the users to collect the reviews for a particular drug
   - Technology used: Python, Angular JS, HTML

2. **Web crawling:** We implemented web crawling to collect drug info, customers reviews and their ratings.
   - Converted Unstructured data to structured data
   - Implemented user interface for data overview
   - Technology used: Python
   - File Format: JSON, CSV

3. **Sentiment Analysis** – We performed various analytics techniques to classify drug effects based on various factors such as age.
   - Text clustering to identify drug effects
   - Algorithm training with keyword for sentiment analysis
   - Analysis of clusters to identify effected age groups
   - Word cloud for visualizing drug effects.

- Technology used: R for word cloud formation and sentiment analysis (classifying word as positive and negative)
  SAS eMiner : Cluster Formation
  SAS eGuide : Used cluster formation as the input and computed mean customer satisfaction level for both the clusters.

4. **Validating the effects**: We validate the effects of drugs with respect to customer age. We Computed Customer satisfaction, Effectiveness Mean, Age Group and then categorized the data in 2 clusters using R word cloud. For both the clusters we are taking the ratings of the customers and calculating its mean which we performed in SAS enterprise guide.

   - Age versus drug effects – we plotted the graph between age as explanatory variable and satisfaction mean as response variable
   - Overall wellbeing with few identified effects – we plotted the graph between age as explanatory variable and satisfaction mean as response variable

5. **Text Analytics using Signals**: By web crawling the data, we fetched the data in JSON format. Then we converted JSON file to CSV and provided as input to Signals for sentiment analysis.

Sources:

1. Webcrawl.py – Its developed for web crawling from WebMD website
2. Cluster_wordcloud.R
3. Sentiment_Analysis.R