**Sentiment analysis on Healthcare Tweets**

Project Proposal By: Mingming Hu

**Introduction & Motivation**

During recent decades, there are a large number of public health problem existing in our life, hence for our project we are trying to explore how machine learning can help us solve these problems. Our main focus would be to discover adverse effects of Claritin with crowdsourcing (tweets on Claritin).

Claritin is a combination of antihistamine and decongestant used to treat allergies, nasal congestion, and [sinus](https://www.rxlist.com/script/main/art.asp?articlekey=22702) pressure. Claritin is available over-the-counter and in [generic](https://www.rxlist.com/script/main/art.asp?articlekey=33074) versions. Claritin’s top 10 adverse events reported by FAERS to FDA include dizziness, drowsiness etc, no of events reported are not high enough to consider Claritin as a health concern.

FAERS called FDA Adverse Event Reporting System, collects mandatory adverse event reports from drug manufacturers, and other medical professionals & consumers. In past drugs like Vioxx and fen-phen were withdrawn due to safety concerns as a result of Serious Adverse Events (SAEs) that were reported to the FDA through FAERS. Hence it is important to constantly verify the data reported to FAERS is correct.

**Overview of Data Set**

Datasets include tweets (on Twitter) that contained ‘Claritin’ for the month of October 2012. Contributors of data-set *Dave Oleson & Crowd-Flowe*r have applied some basic filtering for spam and this has left us with approximately 4,900 tweets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate, Text | **Number of Instances:** | 4900 | **Area:** | Medicine, Healthcare |
| **Attribute Characteristics:** | Categorical,  Integer | **Number of Attributes:** | 17 | **Date Donated:** | Nov 13,  2013 |
| **Associated Tasks:** | Classification, Clustering | **Missing Values:** | yes | **Techniques:** | NLP, Neural Network, SVM |

Attribute Details: Categorical attributes include dizziness, convulsions, heart palpitations, shortness of breath, headache, drug effect decreased, allergies worse after taking a drug, bad interaction between Claritin and other drugs, nausea, insomnia. Demographic information includes ID, gender. Column such as relevance and  sentiment would serve as labels.

We also have data reported to FDA for top 10 adverse effect of Claritin. Based on these data from these two sources we want to perform sentiment analysis on the Claritin effects.

**Problem Statement**

We hope to build a Machine Learning model that can forecast Sentiment score of the Twitter tweet.

1. **Objective**: Assign a Sentiment Score to the Claritin Tweet
2. **Hypothesis**: Total negative score of Claritin Tweets outperform the Side effects reported by drug producing companies to FDA

**Models & Technology**

According to the data characters and the model that we choose, the Natural Language Processing (NLP) will be applied.

Based on NLP machine learning algorithm, we will focus on the following steps - *Tokenization, Stemming, Lemmatization, Punctuation, Character count, Word count, Text classification and Deep learning for NLP,* etc

1. **Visualization Libraries:** Matplotlib, Seaborn
2. **Model Libraries:** SCIKIT-LEARN, TensorFLow, Stanford CoreNLP, spaCy

*REFERENCES :*

[*https://www.figure-eight.com/data-for-everyone/*](https://www.figure-eight.com/data-for-everyone/)

[*https://www.figure-eight.com/discovering-drug-side-effects-with-crowdsourcing/*](https://www.figure-eight.com/discovering-drug-side-effects-with-crowdsourcing/)

[*https://machinelearnings.co/text-classification-using-neural-networks-f5cd7b8765c6*](https://machinelearnings.co/text-classification-using-neural-networks-f5cd7b8765c6)