Text Mining Project

DS745 – Spring 2019

By: Deane Buckingham

2019

**Table of Contents**

Dataset description………………………………………………………………………………………………………………………1

Thesis……………………………………………………………………………………………………………………………………………1

Analysis………………………………………………………………………………………………………………………………………2-4

Resources……………………………………………………………………………………………………………………………………..4

**Dataset description**

I chose to analyze the “Drug Review Dataset” hosted on the UCI Machine Learning Repository. Specifically, I ran my model on the “druglibTrain\_raw.tsv” datafile. I first converted it to .csv format and loaded it into ‘R’.

There are 3107 entries in the dataset. Attributes include some basic information such as entry ID (unique identifier), drug name, and the condition the drug is meant to treat.

There are three user-supplied rating attributes; drug rating (scale of 1-10), effectiveness (scale of 5), and side effects (scale of 5).

And finally, there are three review attributes containing user-supplied text; benefits review, side effects review, and comments review.

**Thesis**

Does a drug's rating correspond to sentiment analysis performed on the drug’s comments?

There were many avenues to investigate potential correlations, but I decided to see if a simple and direct correlation could be found between the sentiment of a drug’s comments review and its rating. This analysis will see if the “commentsReview” column is a good estimator of a drug’s rating. If so, the information in this column could be an important indicator and could be further mined for more detailed information as to what matters most to patients in the drugs they use.

**Analysis**

A screenshot of a cell phone

Description automatically generatedI first looked over the “rating” column in the dataset and found that we mostly have high-rated drugs in the dataset. Below is a count of each rating level and a plot to illustrate this. I am hoping that my sentiment analysis will roughly mirror this trend.

I decided to approach this analysis using the “tidy” methodology as outlined by Robinson and Silge. I took the “commentsReview” column and created a tibble from it where each row is one word. This tibble still saved the original instance the word came from in the “int” column. For example. The first 5 rows of the tibble can be seen below. The 1s mean that all these words came from the first instance in the original dataset.

# A tibble: 6 x 2

line word

<int> <chr>

1 1 monitor

2 1 blood

3 1 pressure

4 1 weight

5 1 and

After removing stop words from the tibble, I began the sentiment analysis. I decided to use the “afinn” lexicon since it already scores the sentiment of words on a 10-point scale. Here I summed up the sentiment score of each word within each review. Below are the first five rows of this new tibble. The first column is the instance identifier (the original row number of the review). The second column is the summed sentiment score for that row.

# A tibble: 1,994 x 2

index sentiment

<int> <int>

1 1 -1

2 2 -3

3 3 -4

4 7 2

5 9 -14

You will notice that some instance identifiers are missing. That is because those reviews did not contain emotionally significant terms. These missing values will be imputed in the next step.

Next, I wanted to join this new “sentiment column” with a selection of my original tibble. This time, I did so by creating a new data frame. After I merged the sentiment of each entry into this new data frame, I filled in the missing row instances with a score of 0 (since the sentiment was either neutral or could not be determined). Below are the first five rows of this new data frame

index ID urlDrugName rating sentiment

1 1 2202 enalapril 4 -1

2 2 3117 ortho-tri-cyclen 1 -3

3 3 1146 ponstel 10 -4

4 4 3947 prilosec 3 0

5 5 1951 lyrica 2 0

I reordered this data frame by “rating” in descending order and plotted what the distribution of my “sentiment” column may look like now.

A screenshot of a cell phone

Description automatically generated

Next, I plotted how “rating” correlates with “sentiment”. The vast majority of sentiment scores are neutral. In fact, the vast majority had values of zero. This, again, is likely due to most comments not containing emotionally charged words so as to register a sentiment score.

I also noticed that there is a longer tail of more negative comments. This is discouraging since earlier I found that most review scores erred towards the positive side.

A close up of a device

Description automatically generatedNow I will look at how my sentiment analysis compares with rating.

I had hoped to see a general linear trend in this plot where the ratings correspond with sentiment. I would’ve expected to see a general diagonal pattern in the dots from lower left to upper right, but that did not happen. It appears that my sentiment analysis was a poor predictor of rating.

I looked into my analysis with more detail. I read a number of “commentsReview” entries in-full and how they were scored by sentiment. A lot of the true meaning in the comments were lost during the sentiment analysis. Many of the comments were neutral descriptions of the medication. This explains why so many were not scored (and the high number of zero sentiment scores). I also noticed that the lexicon thought that several neutral terms in the comments were slightly negative. I believe this is because people mainly described their medical condition in the comments, and even a dispassionate explanation will use some less-than-enthusiastic terminology. Therefore, the sentiment scores erred towards the negative side.

In the end, the sentiment scores I calculated were inconclusive regarding patients’ overall rating of their drugs.

**Resources**

**Drug Review Dataset**

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.

<https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Druglib.com%29#>

**Technique Source**

Robinson, David and Silge, Julia. “Text Mining with R – a tidy approach”. Updated: 03-23-2019.

<https://www.tidytextmining.com/sentiment.html#most-positive-negative>