**SYRACUSE UNIVERSITY**

**SCHOOL OF INFORMATION TECHNOLOGY**

IST 707: DATA ANALYTICS



DATA ANALYSIS PROJECT:

PATIENTS REVIEWS ON SPECIFIC DRUGS

**Presented by:**

Patrick Carlin                           Richard Dzidzornu

**1.0 Introduction**

Social media is playing a crucial role in this era nearly in every field like education, health, medical sciences, marketing, finance, travel, demographics, etc. With the growth of online social networks, everyone can easily interact and engage through different modes and mediums of communication with each other and get the latest updates of information on different fields. The pharmaceutical industry for example has benefited from this era of social media and technological advancement, by bringing customers together to share a common interest about the numerous pharmaceutical products that are made available to the public.

Through online social networking, the communication is vastly improved and different interest of information is available on internet easily at the open pace. Different kind of information needs to share to chronicle highlights of potential benefits and harms and availability of utilities of certain insights, items, people behaviors, products, etc. One of the important fields is the medical and health sciences to consider social aspects through online discussions, blogs, reviews, and online survey. The health-related content shared through various online feedbacks or reviews contains hidden sentiment patterns that need to be identified and extracted through different sources ranging from statistical analysis and machine learning algorithm.

In this regard, the online mechanism is very popular these days for online shopping, different products through different websites like online purchasing of medicine at door step. After purchases, several websites and blogs offer customers rate their products according to their satisfaction and quality of products and services and also by providing feedback facility by which customer can comment on a particular medicine or on quality of services.

With the context of health and pharmacological sciences, the patients are able to share their reviews and experiences of signs and impact of a medicine so that it can help to rate the medicine according to its usage, cost and chronic effects. Several studies have been carried out to concentrate on the patient’s information and their related matters particularly reviewing patient’s medication data with varying costs and usage of a drug. One of these websites that such reviews about pharmaceutical products are found is the drug.com

**2.0 Analysis and Models**

**2.1 About the Data**

The dataset for our analysis is the Drug Review Dataset (Drug.com). It was data from <https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29> and crawled reviews from online pharmaceutical review sites.

The dataset provides patient reviews on specific drugs along with related conditions and a 10 star patient rating reflecting overall patient satisfaction. Our intention is to study the sentiment analysis of drug experience over multiple facets. That is sentiments learned on specific aspects such as effectiveness and side effects. We also intend to use the dataset in developing different models to be used to predict areas of the dataset as well the correlation between ratings, conditions, drugs and drugs usefulness.

We first started by loading the two datasets(train and test) into R-studio for data exploration, understanding and cleaning of the dataset.

> # check the number of observations and variables for the train and test dataset

> dim(traindata)

[1] 161297      7

> dim(testdata)

[1] 53766     7

We then combined the two datasets to one dataset for easy cleaning and exploration using the rbind() which allows for two datasets to be combined base on rows.

> # combined the two datasets (train and test) for cleaning purposes

> drugDataset <- rbind(traindata, testdata)

>

> dim(drugDataset)

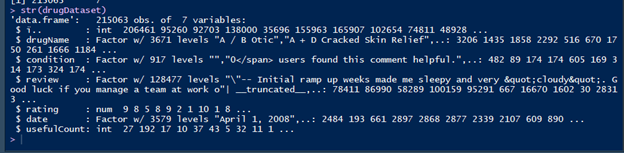
[1] 215063      7

This is the result of looking at the data through the head () command. There are six variables except for the unique ID that identifies the individual, and review is the key variable.

> View(head(drugDataset, 15))>

Below details are additional explanations for variables in the dataset.

* drugName (categorical): name of drug
* condition (categorical): name of condition
* review (text): patient review
* rating (numerical): 10-star patient rating
* date (date): date of review entry
* usefulCount (numerical): number of users who found review useful

The structure of the data is that a patient with a unique ID purchases a drug that meets his condition and writes a review and rating for the drug he/she purchased on the date. Afterwards, if the others read that review and find it helpful, they will click usefulCount, which will add 1 for the variable.

> # fetching the name of all columns/variables

> colnames(drugDataset)

[1] "ï.."   "drugName"   "condition"   "review"   "rating"   "date"   "usefulCount"

First, we will start exploring variables, starting from uniqueID which was initially in the form (‘i..’) as above. We compared the unique number of unique IDs and the length of the train data to see if the same customer has written multiple reviews, and there weren't more than one reviews for one customer.

> # checking the counting of unique numbers in the uinqueID column and comparing it with total count of observations

> length(unique(drugDataset$uniqueID))

[1] 215063

> dim(drugDataset)

[1] 215063      7

This confirms no customer reviewed a drug multiple time.

**3.0 Results**

The ideas of the preliminary data exploration were

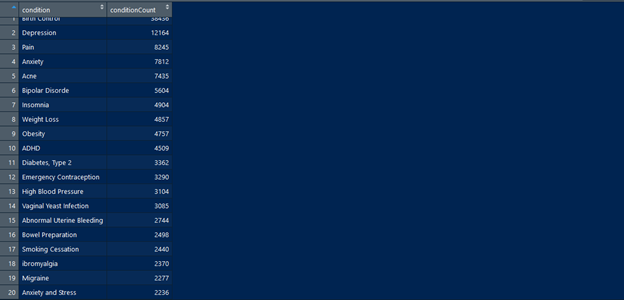
* Most common conditions
* Overall best and worst reviewed drugs
* The curability of each disease
* Best drugs for each condition
* Most useful reviews
* Usefulness vs review score
* Bias in reviews
* Users tend to review things they really liked or really disliked, fewer reviews in the middle

The exploration began with the installation of or calling the r library, the sqldf and ggplot2 packages. A script was developed to generate or fetch the most common conditions in the dataset.

> # using sqldf package/library to extract common conditions and frequency in the dataset

> condition\_SQL <- sqldf("SELECT DISTINCT(condition), count(condition) AS conditionCount +                    FROM drugDataset +                    GROUP BY condition +                    ORDER BY conditionCount DESC")

> View(head(condition\_SQL, n=20))

Conditions like Birth Control, Depression, Pain (in various forms), Anxiety and Acne are among the top five conditions in the drug review dataset.

Since Birth Control happens to be the top condition, further understanding and exploration to the condition was done to have insight of drugs that were taken and reviewed in relation to this condition.

Using the sqldf package, the various drugs used for birth control condition was extracted.

> # exploring birth Control and Drug Name

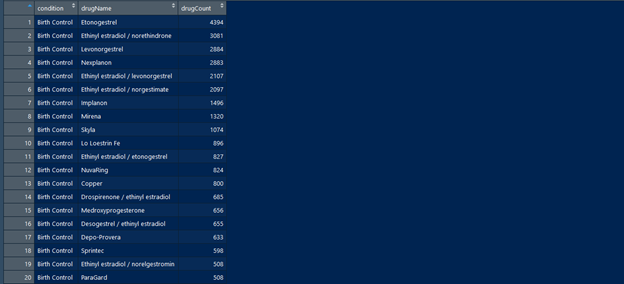
> TestQuery <- sqldf("SELECT DISTINCT(condition), drugName, count(drugName) AS drugCount +                    FROM drugDataset +                    WHERE Condition = 'Birth Control'+                    GROUP BY condition, drugName+                    ORDER BY drugCount DESC")

>

> drugs\_BC <- (TestQuery[1:20,])

>

> View(head(drugs\_BC, n=20))

It can be noticed that common birth control drugs are Etonogestrel, Ethinyl estradiol/norethindrone, Levonorgestrel, Nexplanon and Ethinyl estradiol/ Levonorgestrel are among the top five birth control drugs used and reviewed by customers.

The next area that was explored was top drugs or commonly used drugs and the condition that these drugs relate to.

> # Exploring TOP 20 Drugs per conditions by Customers

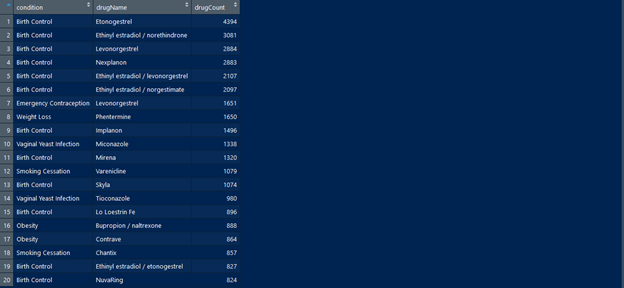
> TestQuery <- sqldf("SELECT DISTINCT(condition), drugName, count(drugName) AS drugCount +                    FROM drugDataset +                    GROUP BY condition, drugName+                    ORDER BY drugCount DESC")

>

> drug\_conditionTOP20 <- head(TestQuery, 20)

>

> View(drug\_conditionTOP20)

Notice the first five common drugs relate to birth control. Other common drugs are Levonorgestrel, Phentermine for weight loss, Miconazole for vaginal yeast infection, and varenicline for smoking cessation.

Rating Distribution

> RatingQuery <- sqldf("SELECT rating, count(rating) ratingCount +                     FROM drugDataset +

                      GROUP BY rating+

                      ORDER BY rating DESC, ratingCount DESC")

> (RatingQuery)

  rating ratingCount

1   10    68005

2    9    36708

3    8    25046

4    7    12547

5    6     8462

6    5    10723

7    4     6671

8    3     8718

9    2     9265

10   1    28918

The highest rating that most of the patients or customers gave for the various drugs was 10 and a rating of 4 being the least rating from customers.

> ################  UsefulCount and Rating Distributing

> # using SQL Query to extract rating distribution

> usefulQuery <- sqldf("SELECT usefulCount, rating+                     FROM drugDataset +                     ORDER BY usefulCount DESC, rating DESC")

> ############       Average UsefulCount vs Rating

> AveUsefulQuery <- sqldf("SELECT rating, AVG(usefulCount) avgUsefulCount +                     FROM drugDataset +                     GROUP BY rating+                     ORDER BY rating DESC")

> (AveUsefulQuery)

   rating avgUsefulCount

1      10       37.47466

2       9       33.78939

3       8       29.29190

4       7       23.199655       6       19.922246       5       17.201537       4       16.525418       3       16.132839       2       16.3363210      1       15.67076

This was graphed to determine whether there is a relationship between ratings and useful counts provided by the customers.

**2.2 Cleaning and Pre-processing**

Based on the understanding attained from the prior data exploration and understanding, the next action taken was to clean the data and process it for analysis.

The column uniqueID was removed from the dataset since it has no significant impact on the genral analysis of the dataset.

Another column, “label”, was added to the dataset. All the Ratings were either labeled as “positive” or “negative” depending on the range between which the rating fell. This was done by using the function cut() to group the ratings into two. Ratings from 1 – 5 were labeled as “negative” and 6 – 10 were labeled as “positive”.

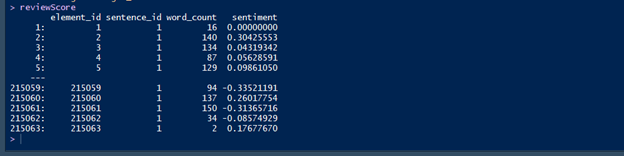
Initial cleaning of the reviews started with removing punctuations and digits. This allowed for sentiment analysis to be performed on the dataset.

The sentimentr package was installed and called from library to be used for this analysis.

A sentiment score was developed using the sentment() function in the sentimentr package.

> # developing sentiment scores for each review

> reviewScore <- sentiment(drugData$review)

The above data depicts the first and last five rows in the dataset with their respective sentiment scores. The sentiment() function for each row determines the score for the reviews base on the count of words in that review. Scare range from –1 to 1. A score of zero(0) being neutral.

Sentiment\_by() function provides a more detailed of sentiment scores in the reviews columns. It allows to create sentiment scores base on averages.

> # create a score base on averages for each review

> aveScores <- sentiment\_by(drugData$review)

Since the object “aveScores” is a data frame, it was saved is a “aveScoresdf” for further process. The negative and positive label was added to the sentiment scores data frame(aveScoresdf). Both the Id and sd columns were removed.

Using the **Naïve Bayes model**

> model\_nbNaive Bayes Classifier for Discrete PredictorsCall:naiveBayes.default(x = X, y = Y, laplace = laplace)A-priori probabilities:Ynegative  positive 0.2988596 0.7011404

Conditional probabilities:

     word\_countY    [,1]     [,2]  negative 82.82742 45.48749  positive 85.66210 45.01827     ave\_sentimentY    [,1]      [,2]

  negative -0.24509008 0.3351107

  positive  0.01033591 0.3313749

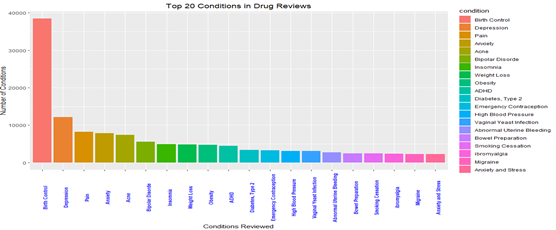
**4.0 Visualization**

Common Conditions

> # visualizing the TOP 20 conditions in the dataset

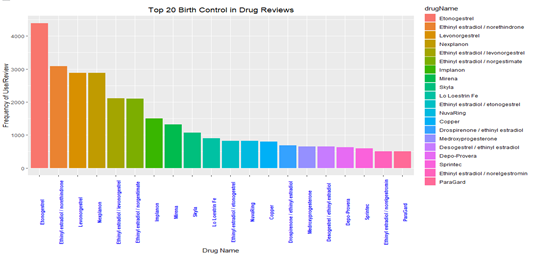
> ggplot(data=commonConditions, aes(x=condition, y=conditionCount, fill = condition)) ++  geom\_bar(position ='stack', stat="identity") + theme(axis.text.x = element\_text(face = "bold", color = "blue", +  size = 8, angle = 90)) + ggtitle('Top 20 Conditions in Drug Reviews') + +  theme(plot.title = element\_text(hjust = 0.5)) + xlab('Conditions Reviewed') + ylab('Number of Conditions')

The graph depicting the top 20 Conditions in the Drug Review Dataset.

Common Birth Control Drugs

> #graph the TOP 20 BIRTH CONTROL drugs distribution by customers> ggplot(drugs\_BC, aes(x=drugName, y=drugCount, fill=drugName)) +  geom\_bar(position ='stack', stat="identity") ++  theme(axis.text.x = element\_text(face = "bold", color = "blue", +                                   size = 8, angle = 90)) + ggtitle('Top 20 Birth Control in Drug Reviews') + +  theme(plot.title = element\_text(hjust = 0.5)) + xlab('Drug Name') +  ylab('Frequency of Use/Review')

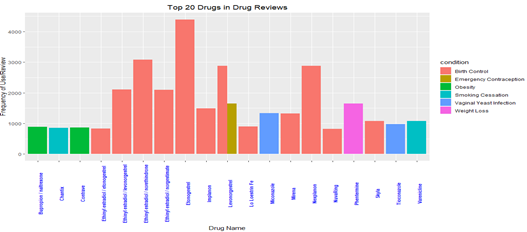
The graph depicting the top 20 birth control drugs in Drug Review Dataset.



> #graph the TOP 20 drugs and Conditions drug distribution of customers

> ggplot(drug\_conditionTOP20, aes(x=drugName, y=drugCount, fill=condition)) + geom\_bar(position ='dodge', stat="identity") ++  theme(axis.text.x = element\_text(face = "bold", color = "blue", +                                   size = 8, angle = 90)) + ggtitle('Top 20 Drugs in Drug Reviews') + +  theme(plot.title = element\_text(hjust = 0.5)) + xlab('Drug Name') +  ylab('Frequency of Use/Review')

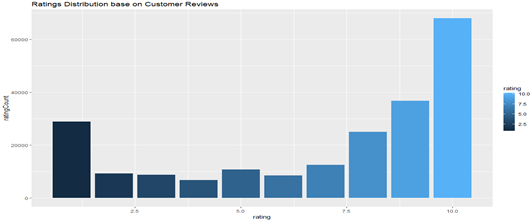
The graph showing the top 20 commonly used drugs in Drug Review Dataset.



> # Graph the rating distribution

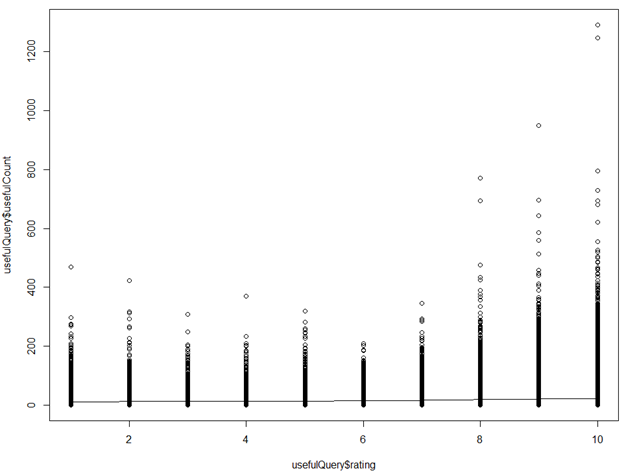
> ggplot(RatingQuery, aes(x=rating, y=ratingCount, fill=rating)) + geom\_bar(stat="identity") ++                      ggtitle('Ratings Distribution base on Customer Reviews') + theme\_classic(base\_size = 10)

The graph showing the Rating distribution base on Customer reviews in Drug Review Dataset.



> scatter.smooth(usefulQuery$rating, usefulQuery$usefulCount)

A Scatter plot showing various ratings and useful counts in drugs in Drug Review Dataset.



> # Graph the usefulCount and ratings distribution using heatmap

> ggplot(usefulQuery, aes(x=rating, y=usefulCount)) + geom\_tile(aes(fill=rating)) +

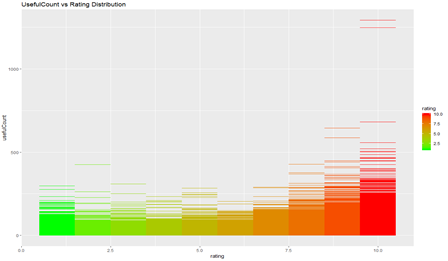
+                        scale\_fill\_gradient(low="green", high="red") +

+                        ggtitle('UsefulCount vs Rating Distribution')

>

Using heatmap to depict the various ratings base on useful counts.

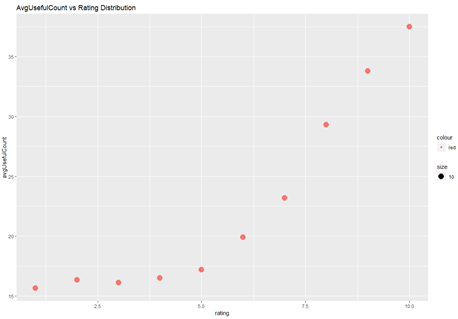
Notice the higher the rating the more it becomes useful to others.



> ggplot(AveUsefulQuery) +  geom\_point(aes(x=rating, y=avgUsefulCount, color='red', size=10)) +

+  ggtitle('AvgUsefulCount vs Rating Distribution')

Noticed there is a linear relationship between customer ratings and the usefulness(usefulcount) to other customers.

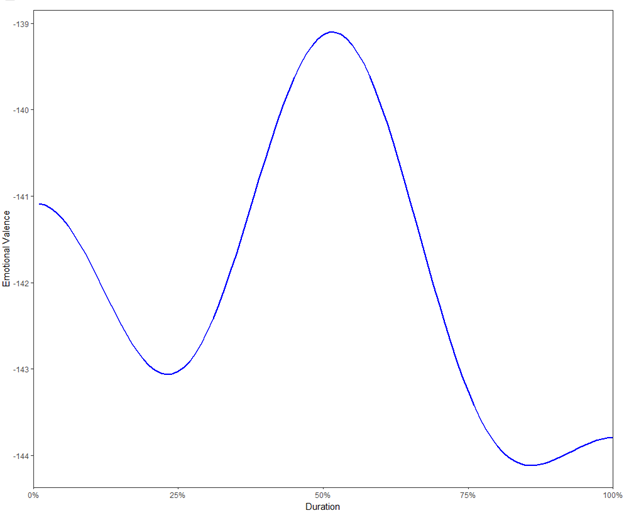


**Sentiment Analysis**

> # Graph the reviews base on the sentiments

> plot(reviewScore)

**The graph depicting the sentiments in the reviews of customers.**

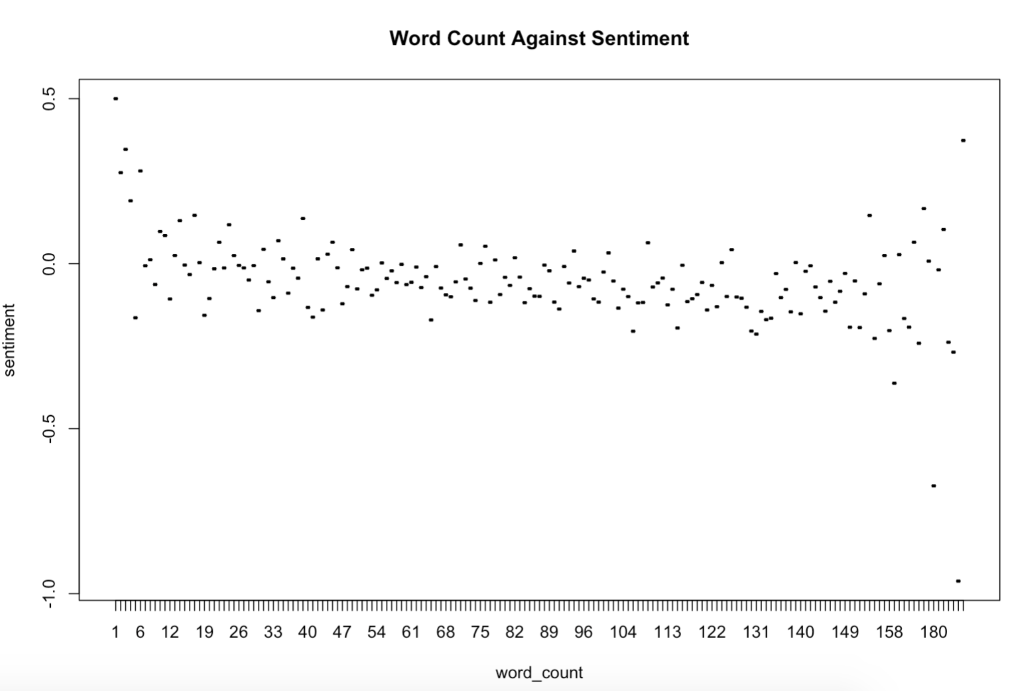


For the drug prediction scoring we wanted to use our sentiment scores to see how and where we would be able to predict what drugs would be appropriate for revision or promotion based on the drugs performance. We wanted to first figure out what would be categorized as “Positive” or “Negative feedback. Giving our sentiment scoring a categorization of “Positive” when it was greater than 0, and negative when it was below 0. Using this we ran through Naïve Bayes, SVM, and Random forest methods to predict the positive and negative outcomes on our sentiment analysis.

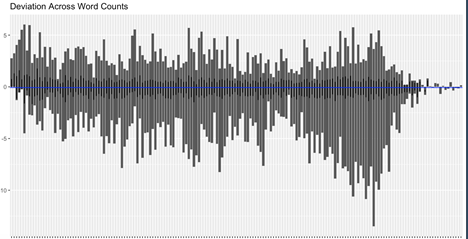
Corpus Word Cloud:



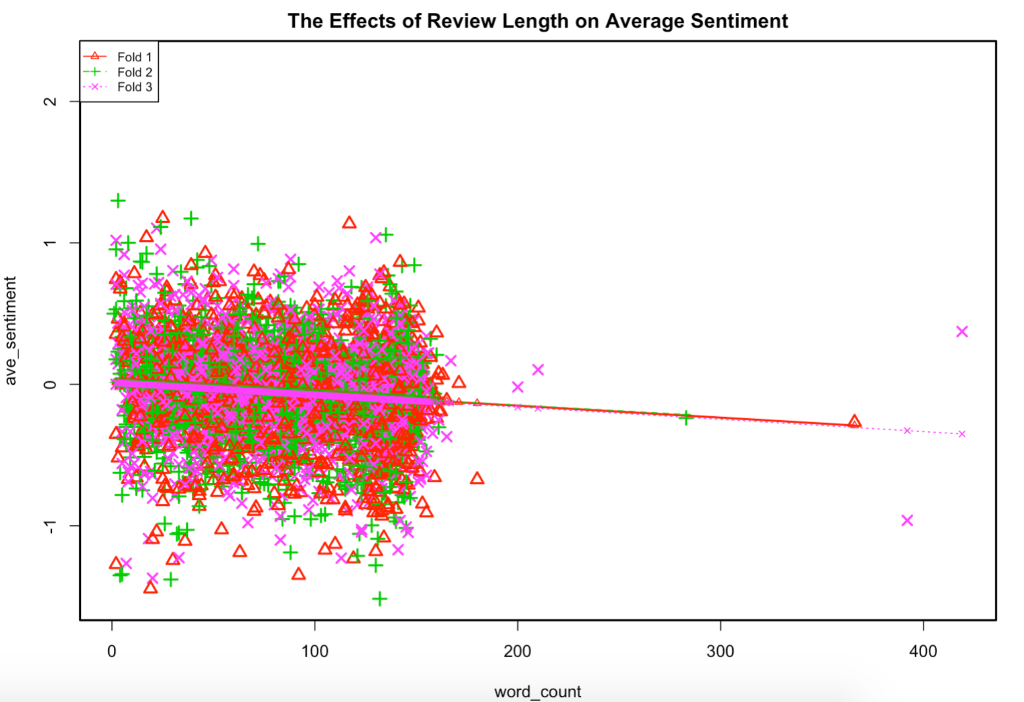
Above is a word cloud of the most used words within our review corpus.



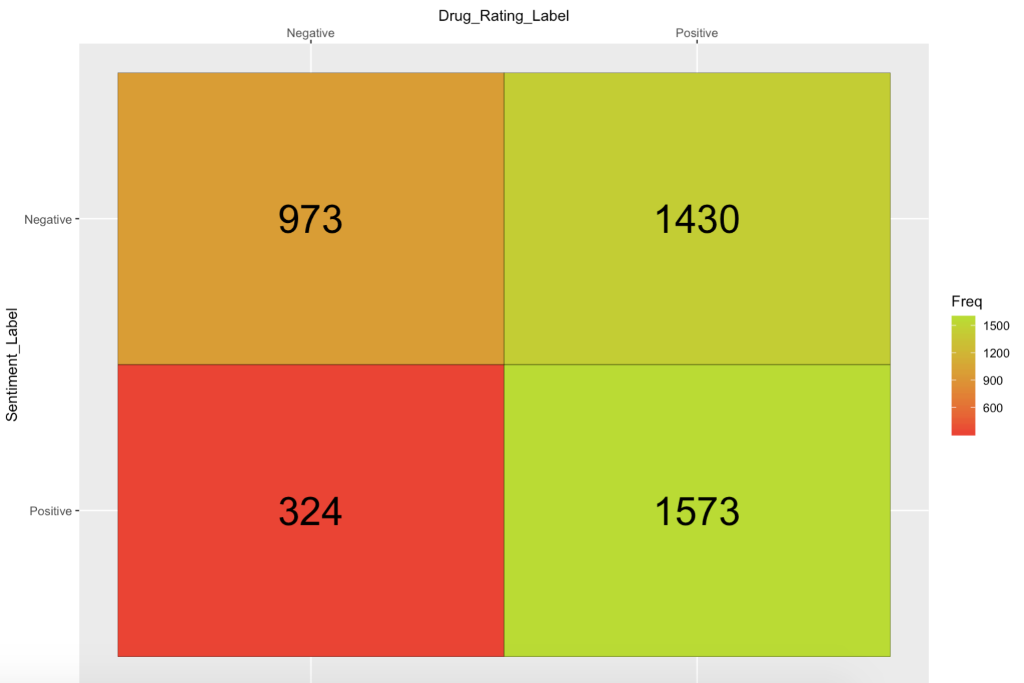
In our first analysis we found that the average sentiment of our drug reviews allows us to see that as work count of a review increases, we found that the average sentiment takes a down turn. We also see that the deviation away from 0 increases around the extreme values. At the beginning we see a more positive sentiment spread. And at around 150 words the deviation spread becomes more negative. To test our hypothesis of the positive and negative sentiment we used a Cross-Validation plot to see how



By summing the average sentiment of each entry, above is a plot of the summed average sentiment for each entry. What we can take away from this article is that there is a higher more positive deviation in reviews with fewer than 50 words. And a higher more negative deviation in reviews that contain more than 150 words.

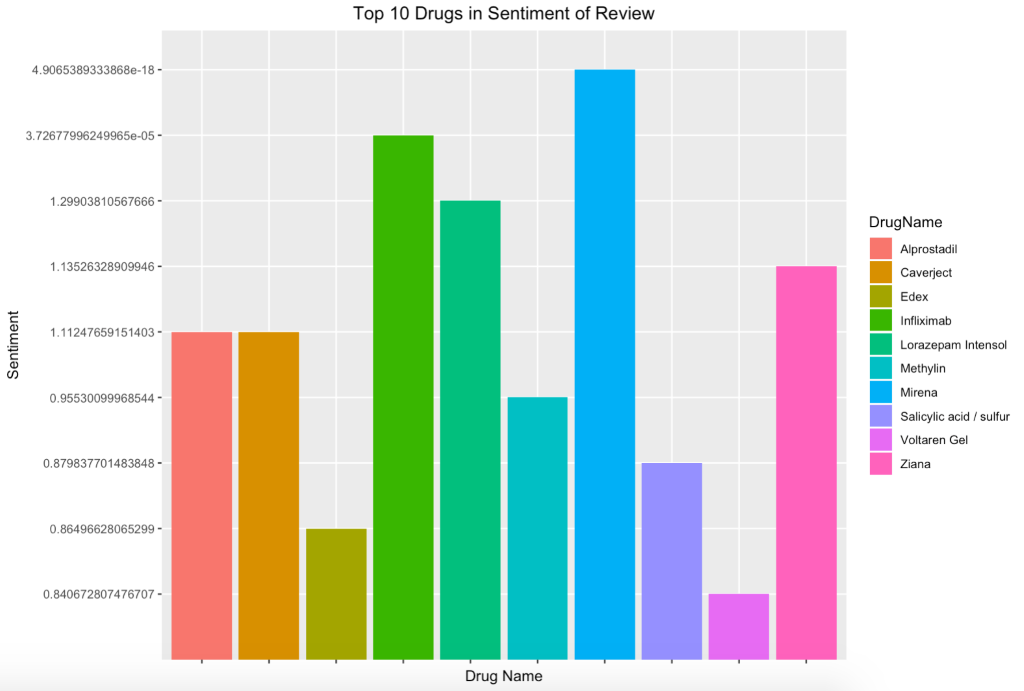


To test our hypothesis of the positive and negative sentiment we used a Cross-Validation plot to see how the sentiment varied over the word count of reviews. There was a very negative slope based on the average sentiment scores within each observed review. By combining this with our sub-set data we can consider the incoming drug reviews that are extremely long would be able to more negative.

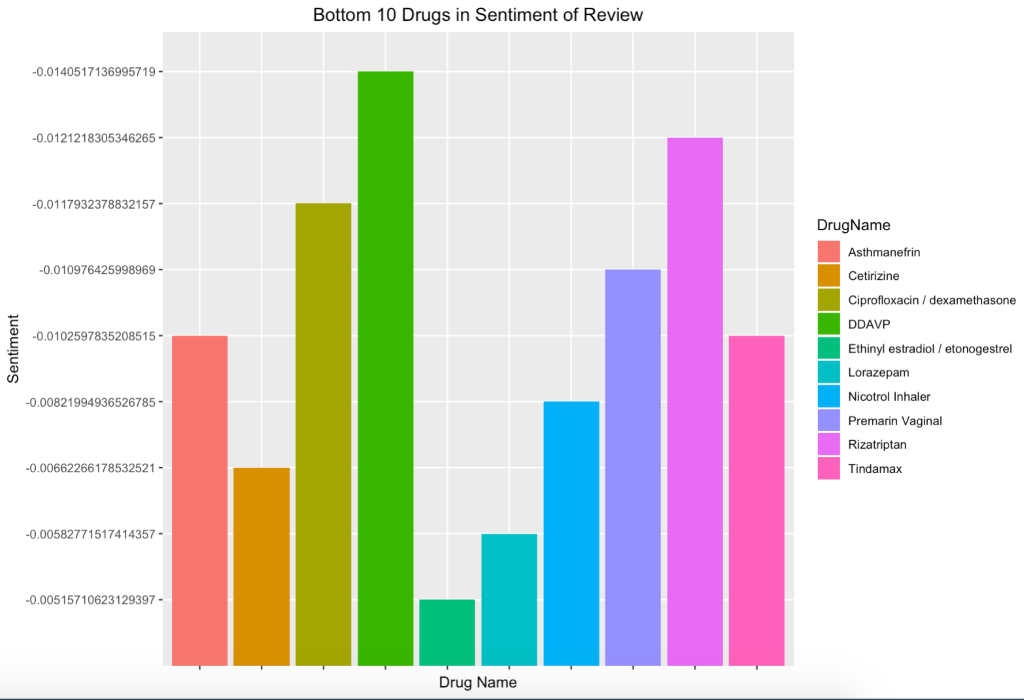


Within our data we were given the actual rating of the drug. So what we wanted to do was to see how well we could predict sentiment based on the Rating label that we created which correctly predicted 59% of the sentiments.

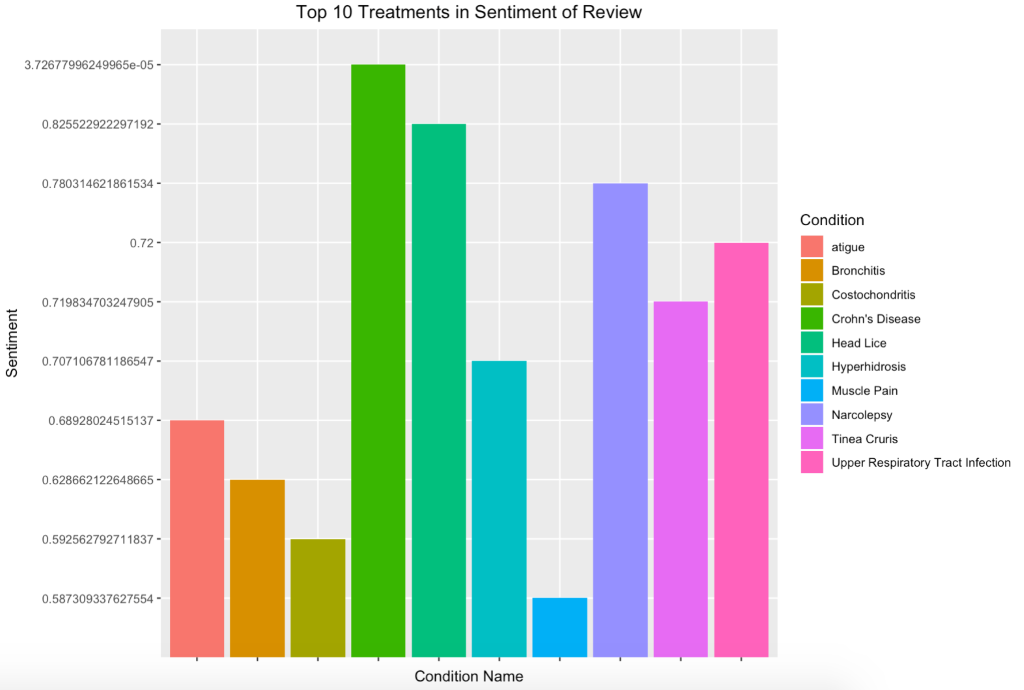
From here the plan was to take it to the business side of things. Given that we had found the top and bottom drugs based on rating, we wanted to predict what we could based on our sentiment levels from the ratings. What we get is the following.



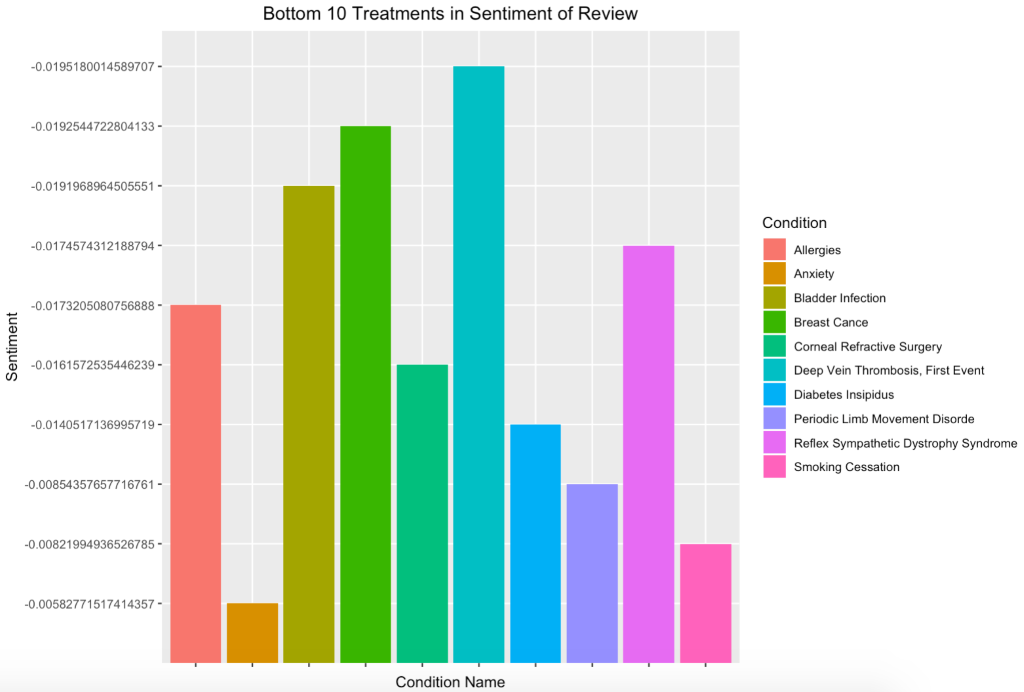
Above is the top 10 drugs based on an aggregated sentiment score. By aggregating our set were able to see that the drugs had very high sentiment scores based off our scoring system. In this case you marketing could narrow its scope down to now specifically focus on these drugs since they have the highest rated reviews.



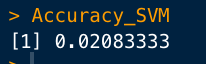
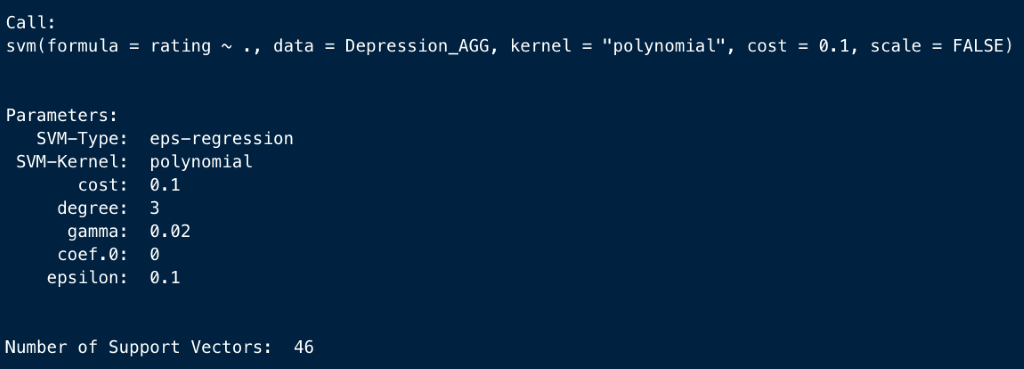
Maybe more interestingly is the graphic with the bottom 10 drugs in drug sentiment. We see that Asthamnefrin, Cetrizine, Ciprofloxacin / dexamethasone, DDAVP, Ethinyl Estradiol, Lorazepam, Nicotrol, Premarin, Rizitriptan, and Tindamax all had the worst aggregated sentiment scores running through our analysis. From a marketing perspective this would be valuable to prove what drugs are the least approved by people. Using the actual review sentiment you would be able to pick out what drugs that need to either be modified or dropped.



Where are analysis lead us was to now figure out what were the top and bottom conditions that were being treated. The above graphic depicts what conditions that are being treated the best based on their aggregated sentiment score. Chrones Disease, Head Lice and Narcolepsy are among the Top 10 treatments with their diseases. From a marketing perspective you can see how valuable understanding the top performers based on customer review can be to a company. Having great customer reviews could get a drug being picked up by distributors and become a value-add a company.

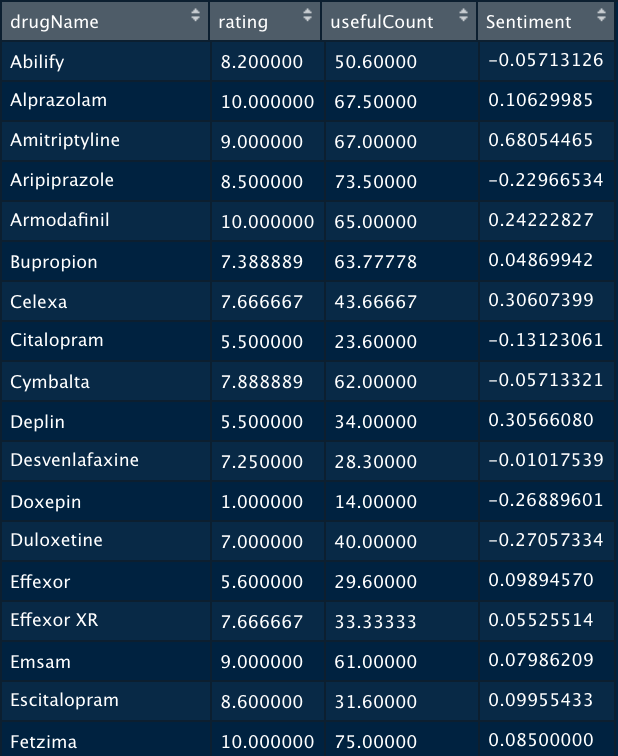


The above graphic is representing the worst 10 treated conditions based on sentiment review. Diabetes Insipididus, Breast Cancer, and Bladder Infections are among the worst reviewed treatments based on our aggregated sentiment review. This is more important than our drug review since we can understand and change the treatment methods. Of course, some of these are issues that can’t be avoided like treatment for breast cancer, but it gives a marketing or pharmaceutical company a benchmark as to what drugs can be worked on.

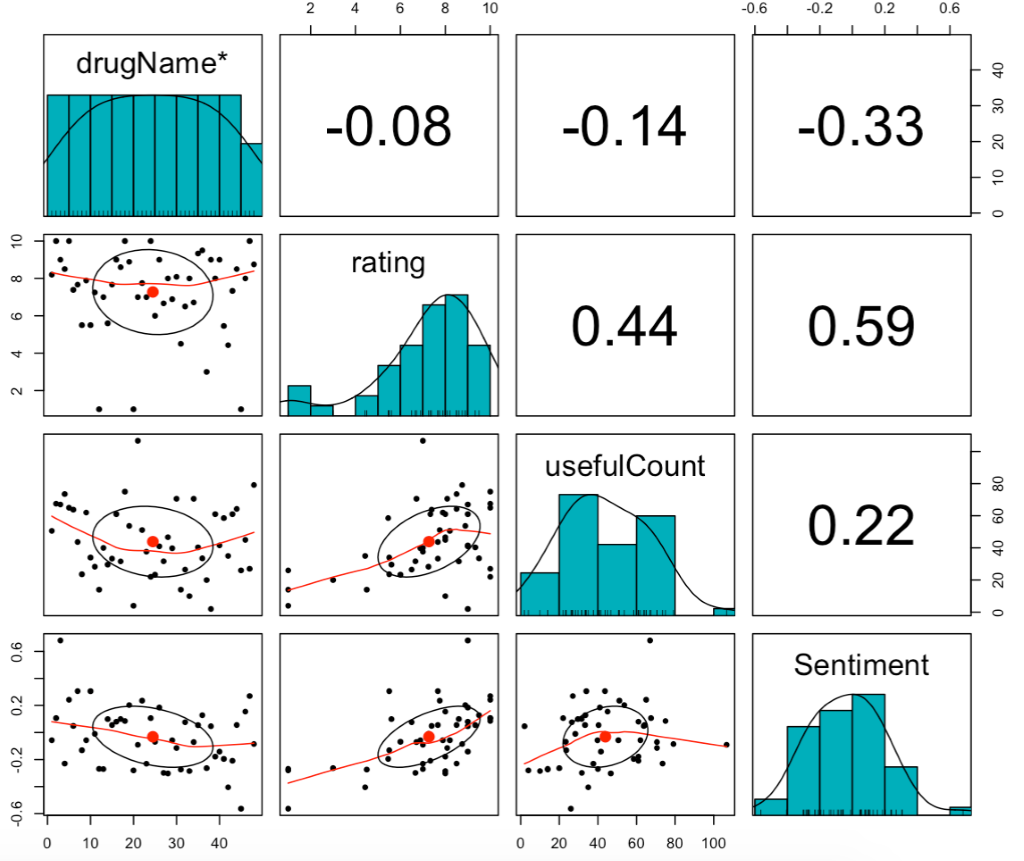


Our initial SVM model we took a grouping of our columns to figure out whether we could successfully predict the rating based on our sentiment score, drugname, condition, and useful count. As you can tell there is little possibility of us accepting the accuracy of around 2%.

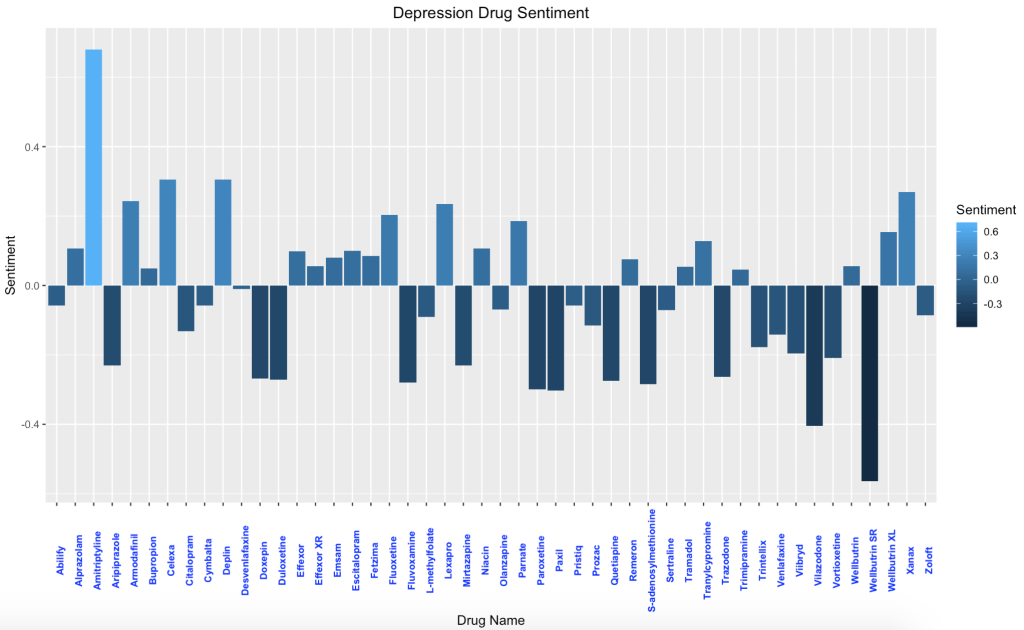
From there we wanted to see if we could dive deeper into one of these categories and figure out what exactly we needed to look at from a marketing perspective. We wanted to figure out what specific drugs are working. For a model we decided to look into all of the drugs used to treat Depression. By aggregating all of the different drugs and instances that account for the treatment of depression these are the average ‘Rating’ ‘UsefulCount’ and ‘Sentiment of the different drugs within our data.



By combining all the entries for the different conditions, we can pinpoint specific values for each drug being used. You can see the different columns, all being referenced to treat Depression, average rating, average useful count, and average sentiment for each aggregated drug.



Using our aggregated dataset that contained specific treatment drugs for depression, above is how the correlation of the variables impacted one another. Usefull count and sentiment had a positive correlation of .22. Sentiment and rating had a positive .59 correlation, so on and so forth.



Above is the graphic that we really would like to focus on. This is a breakdown of how each sentiment was dispersed for each of the different drugs that are treating depression. Very interesting as we see a couple are considered outliers with Wellbutrin SR has an average sentiment of review of around –0.45. One recommendation that we would have to have with this would be to work on those drugs or just drop them from being used.

**4.0 Conclusion**

In the analysis of our data, it was noticed that no customer provided more than one review. In line with that, the most common conditions whose drugs were reviewed relate to Birth Control, Depression, Pain, Acne, Anxiety, Bipolar Disorders, and Insomnia.

Based on the sentiment placement for each drug for treatment it is safe to use the prediction method for customer preference. By analyzing the sentiment scores for the drug reviews, we can see the breakdown of each drug for a treatment, some with a high score for the treatment, and some with low scores for the treatment. You can see consumer preference which allow pharmaceutical companies to adjust defective drugs that get low review scores.

It was also evident that the useful count of a review is determined by the rating of the drug. Thus, there a linear relationship between useful counts and drug ratings.  Most reviews that expressed positive sentiments were of high ratings and reviews with low rating were more less had negative sentiment score.