

Assignment 4: Sentiment Analysis Using DB2 Warehouse on Cloud

Daanish Ahmed

DATA 650 9041

Spring 2018

dsahmed2334@yahoo.com

Professor Elena Gortcheva

UMUC

April 22, 2018

Introduction

Sentiment analysis is an important development within the field of natural language processing (NLP). It involves studying emotions, opinions, and attitudes within a document to understand the author's sentiment towards a certain person, organization, or subject (Cambria, Das, Bandyopadhyay, & Feraco, 2017). Sentiment analysis has many applications—it can be used to determine customer opinion regarding a certain product or to evaluate the public's reaction to a political event. In this assignment, I will perform sentiment analysis on a Claritin side effects dataset using IBM's DB2 Warehouse on Cloud service. The dataset consists of 4900 Twitter tweets describing some of the common side effects that users have experienced after taking Claritin (Oleson, 2013b). My goal is to analyze patient sentiment and determine how these side effects impact user opinion towards this medicine. By finding the relationship between side effects and overall sentiment, our organization can identify the causes of user dissatisfaction and work towards improving the medicine. Within DB2 Warehouse on Cloud, I will use SQL to explore the data and R to build predictive models and visualizations. My methods will include a pie chart, bar graphs, logistic regression, word cloud, and both k-means and hierarchical clustering.

This dataset contains 4900 observations (tweets) and 17 variables. Some of the variables include tweet content, time, sentiment, gender, dizziness, convulsions, shortness of breath, headaches, nausea, insomnia, and bad interactions with another drug (Oleson, 2013a). Sentiment will be the target for most of my models. It is a numeric variable with values from 1 to 5—representing very negative to very positive sentiment (Oleson, 2013b). All remaining variables are in character format. Each of the side effect variables (such as dizziness, allergies, and insomnia) contain two possible values: “yes” meaning that the patient is experiencing that symptom, and “no” meaning that the patient does not have the symptom.

Data Loading and Verification

The first component of the analysis is to load the data into DB2 Warehouse on Cloud. I first selected my dataset, “ClaritinSideEffects.csv,” as the source. Next, I created a new table named “Claritin” within my schema and designated this table as the target into which I would load the data. After this step, I renamed some of the variables to make the SQL and R coding easier. Many of these variables had very long names, such as “bad interaction between Claritin and another drug” and “caused insomnia (the person wasn’t able to sleep).” These two variables were renamed to “bad_inter” and “insomnia” respectively. Other name changes include renaming “heart palpitations” to “heart_palp” and “shortness of breathe” to “short_breath.”

After this, I finalized the load and verified that the data was loaded correctly. According to the load status page, all 4900 rows were loaded successfully and there were no errors or warnings (see Figure 1 in Appendix A). This figure also shows that the data was loaded into the Claritin table on April 11, 2018 at 7:19 PM and that it took approximately 1 second to load. There were several additional steps that I took to verify that the data was loaded correctly. First, I examined the table itself to check that the rows show up properly and that the variables are of the correct type (see Figure 2 in Appendix A). Based on this figure, the data table appears as expected. Next, I examined the row counts in both SQL and R. By viewing the total number of rows in SQL, I found that there are exactly 4900 rows (see Figure 1 in Appendix B). In R, I used both the “nrow” command and the SQL “SELECT” query to verify the row count (see Figure 1 in Appendix C). Both commands produced the same result of 4900 rows, which suggests that all observations were successfully loaded. The final verification step is to confirm that the Claritin table exists in R by using the “idaExistTable” command (see Figure 2 in Appendix C). Based on the output, it is clear that the Claritin data has been successfully loaded into the server.

SQL Methods

Once the data has been successfully loaded, I can perform an exploratory data analysis on the table using SQL. All of these queries can be found within the attached SQL script. My first query—aside from displaying the row count—involves showing the number of rows for each non-null sentiment level (see Figure 2 in Appendix B). This query will provide us with a general understanding of how most users feel about Claritin. According to this figure, the vast majority of users have a sentiment level of 3 or 4, with 1421 and 2328 rows respectively. This means that most users have a relatively favorable view of Claritin. However, there are still about 700 people altogether with unfavorable sentiment levels of 1 or 2.

The next query analyzes the number of tweets from male and female patients, as well as the average sentiment for each gender (see Figure 3 in Appendix B). It is useful for determining whether a patient's gender will affect their sentiment towards Claritin. The average sentiment was obtained by using the "AVG" function in SQL. Based on the results, there are almost 1000 more tweets from females than from males—with 2557 female users and 1558 male users. This may imply that women are more likely to tweet about their medical condition than men. Furthermore, this figure also reveals that the average sentiment for women is 3.37, while the average sentiment for men is 3.49. Women have slightly less favorable opinions than men regarding Claritin—which may suggest that female patients are more likely to suffer from side effects after taking the medicine. At the very least, they may be more likely to report the symptoms, or their symptoms may be more severe than on male patients. But this is only a theory, since the sample size of male patients is significantly smaller than that of female patients.

My third query displays the number of rows per sentiment level for users with any of the possible side effects (see Figure 4 in Appendix B). It functions similarly to my first query, but it

only includes patients suffering from at least one of the ten side effects. Its purpose is to highlight the impact of side effects on the overall user satisfaction score. The output shows that the vast majority of patients have low sentiment levels of 1 or 2—with 185 having a sentiment of 2 and 50 having a sentiment of 1. These results contrast sharply with my initial query, in which most patients have sentiment levels of 3 or 4. These findings are not surprising, since they indicate that having any negative symptoms will cause patients to think very poorly of Claritin.

Next, I examined the number of rows and average sentiment of users without any side effects (see Figure 5 in Appendix B)—as well as those with at least one possible side effect (see Figure 6 in Appendix B). These queries show how often side effects occur and how severely they affect user sentiment. These figures indicate that only a small minority of the sample population has experienced these side effects, since only 286 out of 4900 people reported having any symptoms. Nevertheless, the average sentiment declines massively from 3.49 for those with no symptoms to only 2.08 for those with at least one symptom. This means it is unlikely that a patient will experience any negative effects after taking Claritin. But if they do, they will most likely associate the symptoms with Claritin itself—hurting the company’s reputation.

From here, I created a query that examines the average sentiment for patients with any of the following five symptoms: headaches, shortness of breath, worsened allergies, nausea, or insomnia (see Figure 7 in Appendix B). Its purpose is to determine whether certain side effects have a greater impact on user sentiment. Based on the results, all of these symptoms lead to a very unfavorable sentiment level close to 2. Insomnia has the highest average sentiment of 2.38, while having worse allergies will produce the lowest average sentiment of 1.73. Since allergies appear to impact user sentiment the most, it is likely that allergic reactions caused by Claritin are more severe than the other included side effects. However, not all side effects occur with the same

frequency, meaning that the information for some side effects may be insufficient. But since allergies produce the lowest sentiment, I will examine this side effect in greater detail.

My next query shows the count and percentage of rows in each sentiment level for patients with worsened allergies after taking Claritin (see Figure 8 in Appendix B). Its purpose is to analyze the impact of this specific symptom on the user sentiment levels. According to this figure, 64.4% of all patients with worse allergies have a sentiment level of 2, while 31.8% have a sentiment level of 1. Only about 3.8% of users with allergies have positive (or even neutral) sentiments of at least 3. As such, these results show that nearly all patients suffering from worse allergies after taking Claritin will have highly unfavorable opinions about the medicine.

My final query examines the tweet content, gender, and sentiment level for all patients with headaches after taking Claritin (see Figure 9 in Appendix B). The reason why I am examining the tweet content is because it is critical to analyze the text itself to gain insights regarding the possible sentiments. This query returns 7 rows involving 4 female patients, 2 male patients, and one user with an unspecified gender. Of these 7 patients, two have sentiment levels of 3 while all others have sentiment levels of 2. There are only 7 rows, meaning that it is very rare for headaches to be caused by taking Claritin. Although the sample size is small, the results imply that female patients are much more likely than male patients to report these symptoms. Furthermore, only two users have neutral sentiment levels of 3—with all others having negative sentiments of 2. When examining the tweet content, we find that most users believe Claritin itself causes their headaches. For instance, one patient writes “now I remember why I don’t take Claritin” while using the hashtag #massiveheadache (Oleson, 2013a). Overall, these queries suggest that most users with negative symptoms will have very negative feelings about Claritin. But these cases are rare, since the vast majority of Claritin users report having no side effects.

Data Preparation Using R

With the exploratory data analysis complete, I will now prepare the data for analysis using R. All of my R code is found within the attached R script. The first step is to connect to the DB2 Warehouse database server and load the data. This step requires installing and loading the `ibm_dbR` package (“In Memory Analytics,” 2018). Next, I ran the code to access the server, which requires entering user credentials such as host name, username, and password. These credentials can be accessed from the DB2 Warehouse on Cloud “service credentials” page. After accessing the database, I loaded the data from the Claritin table into a new data frame “CLAR_SE.”

From here, I performed several validation steps to ensure that the data was loaded into R correctly. Two of these steps were discussed earlier in the paper. Namely, I verified the row count in both the data frame and the database table (see Figure 1 in Appendix C). I found that both commands returned the same output of 4900 rows. Secondly, I used the “`idaExistTable`” command to ensure that the Claritin table exists in the database (see Figure 2 in Appendix C). Afterwards, I performed one final validation step by comparing the number of tweets per sentiment level in both the data frame and the database table (see Figure 3 in Appendix C). The output shows that both data sources contain the same number of rows in their respective sentiment levels. For instance, both sources have 191 cases with a sentiment of 5, 2328 cases with a sentiment of 4, and so on. Additionally, these numbers are identical to the sentiment counts in my original SQL query (see Figure 2 in Appendix B). Based on these findings, the data has been loaded correctly.

After validating the data, I can proceed to the data preprocessing stage. The first step is to initialize the packages required for this analysis. I installed the “`tm`” and “`SnowballC`” packages for text mining and the “`wordcloud`” package for generating word clouds. I also activated “`ggplot2`” for creating bar plots and “`cluster`” for k-means clustering. Next, I removed variables

that are not useful for the analysis. The interaction ID and article URL are unique values with no analytical value, so I set these variables to null. Afterwards, I examined the data to check for missing values. I found that “sentiment,” along with all “side effects” variables (such as dizziness and headaches) contained exactly 232 missing values (see Figure 4 in Appendix C). By displaying the rows containing missing values, it is evident that all cases of missing values are within the same 232 rows (see Figure 5 in Appendix C). Therefore, I removed these rows from the data frame. After this, I verified that all missing values have been removed.

The next step is to convert all character variables into factors except for “content” and “time.” Most of the variables are categorical, and converting them into factors will help when making visualizations such as pie charts and bar graphs. I verified that the variables were converted into factors by examining the descriptive statistics (see Figure 6 in Appendix C). This figure also shows that some additional preprocessing steps that will be needed. Firstly, the “relevant” variable indicates that some tweets are not written in English. Since I will perform text mining on the data, it is necessary for all documents to be written in the same language. Thus, I removed all rows with non-English tweets. Furthermore, gender has 274 rows with “bad link or company” information. To ensure that the data is complete and without errors, I removed the rows with bad gender information. After these steps, I found that “relevant” and “convulsions” each contained only one level. These variables are no longer necessary, so I removed them.

From here, I created a new variable named “symptoms” that lists the name of the side effect (if any) that each patient has. The reason why I created this variable is because it will help for certain visualizations for which I compare the frequencies of each side effect. It allows me to analyze the frequencies of all side effects at the same time, instead of examining each side effect separately. This task is much easier if all symptoms are included in the same variable instead of

being listed as separate variables. The main shortcoming of this approach is that if a patient is suffering from more than one side effect, the variable will still only list one side effect. After creating this variable, I converted it into a factor and verified that it was created correctly (see Figure 7 in Appendix C). According to this figure, allergies are by far the most common side effect with 127 cases. However, the vast majority of users (4118) have not suffered from any side effects after taking Claritin. With this step completed, the data is ready for analysis.

R Methods

I will now describe the methods that I will use in my analysis. I will first create a pie chart that shows the frequency of each possible side effect. Each symptom will be given a unique color. This graph is useful because it allows us to determine which side effects are the most common. This visualization requires creating a copy of the Claritin data frame that excludes all patients with no symptoms. This is because 4118 users have experienced no side effects, and failure to remove these rows will make it very difficult to see the cases with side effects on the graph.

Next, I will create two bar graphs—the first of which shows the number of patients with each sentiment level labeled by their symptoms. It is useful since it shows which side effects are the most common in each sentiment level. The second bar graph will show the number of patients with each side effect, labeled by gender. It will help us to determine whether males or females suffer the most from certain symptoms. These bar graphs will be created using the “ggplot” command, which requires the “ggplot2” package in R. Furthermore, both graphs use the data frame that excludes patients with no symptoms. This allows for the graphs to highlight each individual side effect, instead of being dominated by cases involving no side effects.

After creating these initial frequency plots, I will build a logistic regression model to predict the patient's sentiment. My goal is to predict whether patients will have a negative sentiment (1 or 2) or a non-negative sentiment (3, 4, or 5). Before creating this model, I made a copy of the original data frame that excludes all unnecessary variables. Tweet content, time, and "symptoms" will not be used in this model. Next, I restructured the target variable "sentiment" into a binary factor variable with two levels: '1' indicating a negative sentiment of 1 or 2, and '0' indicating a positive or neutral sentiment of 3, 4, or 5. This step is necessary because logistic regression models generally function better on binary targets. To create the model, I will partition 70% of the data into the training set and 30% of data into the test set, using a random seed to reproduce the results. Next, I will build the model on the training set with "sentiment" as the target and all remaining variables as inputs. To evaluate the model, I will create confusion matrices for both the training and test data. These will allow us to compute the classification accuracy, as well as the number of true positives (sensitivity). Ideally, we want our model not only to have a high accuracy, but also for it to effectively classify true positives. If the model performs much worse on the test set than on the training set, we can conclude that overfitting took place.

After logistic regression, I will perform several text mining operations to analyze the content of the tweets. My focus will be on tweets with negative sentiment levels of 1 or 2. My methods will include word cloud analysis, k-means clustering, and hierarchical clustering. Before implementing these methods, it is vital to perform text preprocessing on the tweet content. These steps require the "tm" and "SnowballC" packages to be activated. First, I created a data frame that only contains the tweet content where the sentiment value is either 1 or 2. I then stored these tweets into a corpus. By examining one of the tweets, we see that the text needs to be cleaned thoroughly before starting the analysis (see Figure 13 in Appendix C).

To clean the texts, I removed URLs, numbers, punctuation, special characters, and non-ASCII characters. Next, I removed stop words by using the built-in English and “Smart” stop words lists (Feinerer & Hornik, 2017). I also created an additional list of words to remove, which consists of additional stop words not included in the previous lists. This list also removes unwanted terms (such as profanity), and it removes the word “Claritin” because it is trivial and appears in every tweet in the dataset. I then performed stemming to reduce every word to its root word. I removed stop words again after stemming, since some stop words may have been formed by this process. I changed all letters to lowercase and removed the extra whitespace between terms. By examining the same tweet again, we see that it has been cleaned (see Figure 14 in Appendix C). Finally, I used this corpus to create the document term matrix (DTM).

My first text mining operation involves creating a word cloud to showcase the most frequent terms associated with a negative sentiment. It is useful because it shows not only the most widely used words in these tweets, but also how often those words are used. This method requires installing the “wordcloud” package. The first step is to create a list of all unique words and their frequencies. Next, I set up a color scheme that allows up to 6 colors to label words according to their relative frequency counts. I then generated the word cloud with a maximum of 60 words—using a random seed to recreate the results.

The final component of my analysis involves using both k-means clustering and hierarchical clustering to categorize words based on their frequencies. I am using both clustering methods to determine which is more effective at classifying the words in the Claritin dataset. First, I removed sparse terms from the DTM using a threshold of 0.985. Failure to remove sparse terms will result in each clustering model having over 1000 terms. This is a problem because hierarchical clustering works better when using smaller datasets with no more than 150 observations (Bati,

2015). The reason I used a high threshold of 0.985 is because setting it too low will remove too many terms from the DTM. After removing sparse terms, examined the properties of the DTM and found that there are 39 terms remaining (see Figure 16 in Appendix C). This is a good number of documents to use for k-means and hierarchical clustering.

I then created a dissimilarity matrix (DSM) to compute the distances between and within clusters. To determine the ideal number of clusters k , I will first use the elbow method—which involves plotting the within clusters sum-of-squares over k and selecting the approximate k -value where the plot appears to form the shape of an “elbow” (Bati, 2015). However, the elbow method only works if there is a single point that clearly resembles an “elbow.” If there are multiple points that may qualify as such, it is better to use a different method for selecting k (Ng, 2017). If this happens, I will use the equation $k \approx \sqrt{n/2}$, where n is the number of instances (Bati, 2015). Once I find the ideal number of clusters, I will build the k-means clustering model by using the “kmeans” method with the DSM and obtained k -value as inputs. I can then evaluate the terms that appear in each cluster. Finally, I will plot a hierarchical cluster dendrogram with the same DSM as an input. Using the same k -value as well, I will draw red boxes to outline the clusters—allowing comparison between the two clustering models. In the following sections of this paper, I will describe the results of every method used in this analysis.

Pie Chart

I will first describe the output from my pie chart that shows the frequency of each side effect. As mentioned earlier, this pie chart does not contain any users who experienced no side effects. Based on this image, we see that allergies are by far the most common side effect—taking

up almost half of all symptoms listed (see Figure 8 in Appendix C). Interestingly, allergies also had one of the lowest average sentiment levels of any side effect—with an average sentiment of only 1.73 (see Figure 7 in Appendix B). Of course, this may be due to the likelihood that many allergy cases were more severe when compared to other symptoms. However, it is also possible that the large sample size of patients with allergies allowed for a more accurate depiction of user sentiment regarding this symptom. If other side effects were included more frequently in this dataset, their sentiment may be closer to that of allergies. This is especially true for shortness of breath, nausea, heart palpitations, and headaches—since they have the lowest frequencies in the dataset. If each of these side effects appeared more frequently, then we may be able to gain more insight about their severity and its effect on the overall user sentiment.

Bar Graphs

From here, I will examine my two bar graphs—starting with the first one that shows the frequency of each side effect based on the user sentiment level. Both graphs focus on patients with side effects, omitting any users that did not suffer from at least one side effect. According to this graph, most users with side effects have a sentiment level of 2, followed distantly by a sentiment of 1 (see Figure 9 in Appendix C). In both cases, allergies are the most common symptom. And for users with the lowest sentiment of 1, allergies take up the grand majority of cases. This is not surprising when considering my findings from the previous pie chart. Interestingly, allergies do not make up the majority of those with sentiments of 3 or 4. In these cases, having a bad interaction with another drug is the most common side effect. This may suggest that those who experience bad drug interactions do not suffer as severely as those who experience worse allergies after taking Claritin. Another interesting finding is that a very small

number of patients with insomnia have a sentiment level of 5. This may suggest that those patients may not attribute the insomnia directly to Claritin. Additionally, it may mean that although they suffered from insomnia, they still acknowledged that Claritin has been beneficial to them as a whole. Even though some users may occasionally suffer from certain symptoms, the medicine's benefits may outweigh its shortcomings most of the time.

I will now examine the second bar graph that shows the breakdown of patients' gender for each side effect. Based on this graph, we once again see that allergies are by far the most common symptom, followed distantly by a decrease in drug effect (see Figure 10 in Appendix C). But in addition, we see that female patients are many times more likely than male patients to suffer from every single one of these side effects. This relates to one of my earlier findings, in which female users had a lower average sentiment than male users while at the same time appearing much more often in the dataset (see Figure 3 in Appendix B). This bar graph supports the notion that women experience these side effects more often than men. It may seem that these results are due to the larger sample size of female patients—since there are almost 1000 more posts from female users. However, the bar graph suggests that women are more than twice as likely as men to suffer from several of these side effects. This is a much higher ratio than the actual proportion of women to men in the dataset (which is less than 2 to 1). Therefore, it is highly likely that females are more vulnerable to the side effects associated with taking Claritin.

Logistic Regression

With the first two methods complete, I will now analyze the results of my logistic regression model. The purpose of this model is to predict whether a patient will have a negative

sentiment of ‘1’ or a non-negative sentiment of ‘0.’ First, I will examine the confusion matrix for the training data (see Figure 11 in Appendix C). Based on this figure, we see that the model correctly identified 2559 cases of 0 and 146 cases of 1. The model also has 330 false positives and 10 false negatives, which results in a classification accuracy of 88.8%. This is a very accurate model, although there is still some room for improvement. However, the ideal model should also have a high sensitivity—meaning that it should be effective at finding patients with negative sentiments of 1. Since the model has 146 true positives and 10 false negatives, its sensitivity rate is 93.6%—which makes this model extremely effective at identifying 1’s in the training set.

Next, I will evaluate the model’s performance on the test set and compare its results to those from the training set. Based on the classification matrix for the test data, we see that the model accurately classified 1112 true negatives and 69 true positives (see Figure 12 in Appendix C). It also misclassified 152 false positives and 12 false negatives—which leads to a classification accuracy of 87.8%. Due to its high accuracy rate, this logistic regression model is very effective at classifying user sentiment in the Claritin dataset. And since the test accuracy is only marginally smaller than the training accuracy, it is evident that the model does not suffer from overfitting. When evaluating the sensitivity, we see that the model has 69 true positives and 12 false negatives, leading to a sensitivity rate of 85.2%. This figure is still quite high and suggests that the model is relatively effective at classifying true positives. But it is clearly lower than the training sensitivity of 93.6%, which means that the model may benefit from improvement. One suggestion for enhancing the model is to remove unnecessary variables. This can be done by evaluating the p-values—for which variables with p-values under 0.05 are significant, while those with p-values greater than or equal to 0.05 can be removed (Knode, 2016). By including only relevant inputs, the model’s accuracy and sensitivity for test data will likely improve.

Word Cloud

The next method to analyze is my word cloud, which shows the 60 most frequent words in tweets from users with negative sentiments (see Figure 15 in Appendix C). This visualization shows the words that appear frequently within negative tweets, as well as how often each word is approximately used. Based on this figure, we see that the two most frequent terms by far are “allergy” and “work.” The frequent usage of “allergy” suggests that users with negative sentiment towards Claritin are more likely to suffer from allergies while using the medicine. This is consistent with my earlier findings, in which allergies are the most common side effect and are commonly found in patients with the lowest sentiment levels. The high usage of the term “work” may relate to whether the drug is working as expected. A decrease in drug effect is the second most common side effect in the dataset (see Figure 8 in Appendix C), which may explain why the term “work” is used so often. However, the actual meaning of this word is uncertain because we lack the context in which the term is used. One of the issues of analyzing individual word frequencies is that we are unable to extract the context of the document itself.

By looking at some of the smaller words, we can gain further insights about the causes of low user sentiment. For instance, we find that “Zyrtec,” “Allegra,” and “Benadryl” are all included in the 60 most common words. These terms relate to medicines, and they are likely referring to users having a bad interaction between Claritin and another drug. Other commonly used terms include “sneeze,” “sinus,” “nose,” and “cough.” These words likely refer to allergies or another related side effects for which they contribute to lower user sentiment. The appearance of the terms “head” and “headache” suggests that headaches will also hurt user opinion towards Claritin. The term “sleep” likely refers to insomnia, while the words “drowsy,” “breath,” and “cold” indicate the occurrence of other unwanted symptoms such as drowsiness, bad breath, or colds. Overall, we

see that most of these words refer to the known side effects from taking Claritin. Based on this, low user sentiment is most likely caused by the presence of side effect symptoms.

Clustering

My final set of methods involves performing a clustering analysis on the tweets with negative sentiments, for which I will group terms into clusters based on their frequencies. I will perform both k-means clustering and hierarchical clustering, and I will compare their results to decide which approach is more suitable for this problem. To select the number of clusters k , I first used the elbow method (see Figure 17 in Appendix C). This method shows the within clusters sum-of-squares (in blue) and the between clusters sum-of-squares (in black) plotted over k . Based on this figure, there are several points that can qualify as the graph's "elbow." The points $k=2$, $k=6$, $k=9$, and $k=14$ all appear to form the shape of an "elbow." But since there is no clear k -value that can qualify as the "elbow," the elbow method is not useful for choosing k . Instead, I will use the formula $k \approx \sqrt{(n/2)}$, where n is the number of cases. In this analysis, each case represents a tweet—of which there are 39 (see Figure 16 in Appendix C). By plugging in $n=39$, the equation yields an approximate k -value of 4. Thus, I will build my clustering models using 4 clusters.

After building the k-means clustering model, I will evaluate its properties. This model contains 4 clusters containing 2, 4, 20, and 13 words each (see Figure 18 in Appendix C). By examining the cluster visualization, we can analyze the terms that appear in each cluster (see Figure 19 in Appendix C). The first cluster contains only two terms: "work" and "allergy." This cluster is farthest from any other cluster, meaning that these two terms differ significantly from any other terms in the dataset. In my word cloud visualization, I found that the terms "work" and "allergy"

were by far the most common terms—with no other terms having frequencies that large (see Figure 15 in Appendix C). Therefore, the clustering results support my finding that “work” and “allergy” are the most widely-used terms in the Claritin tweets.

The second cluster contains 4 words: “Zyrtec,” “Allegra,” “feel,” and “hour.” These terms are not used nearly as often as “work” and “allergy.” But in the word cloud, they are larger in size when compared to other terms—meaning that they appear more often than most words. The third cluster contains 20 terms, including “sleep,” “sick,” and “cold.” These terms appear less often than those in the second cluster, but more often than those in the last cluster. The fourth cluster contains 13 terms, which include “sneeze,” “drowsy,” and “sinus.” These terms appear the least often of any words in the current DTM, but they still have higher frequencies than the words that were omitted from this list. One important observation is that there is no overlap between clusters. K-means clustering should not involve overlapping clusters, and every instance must only belong in one cluster (James, Witten, Hastie, & Tibshirani, 2013). Since this condition is true, k-means clustering appears to be a good fit for the Claritin tweets data.

Finally, I will evaluate the dendrogram that I created using hierarchical clustering. Although this method does not require choosing the number of clusters k , I used $k=4$ clusters to allow for comparison between the two clustering methods. Based on the dendrogram, we see that there are two clusters containing one term each, one cluster with two terms, and a cluster that contains 35 words (see Figure 20 in Appendix C). The clusters appear to be grouped according to frequency, with “allergies” and “work” having the highest hierarchical ranking of any terms. Allegra and Zyrtec are grouped into the same cluster, while all remaining terms are grouped into a single massive cluster. But overall, the clusters do not appear to be grouped in the most effective way. For instance, “allergies” and “work” have very similar frequencies—yet they are grouped as

two individual clusters. Furthermore, the terms “Zyrtec” and “feel” have similar frequencies and would ideally appear in the same cluster. However, “Zyrtec” is instead grouped with “Allegra,” which has a noticeably lower frequency based on the word cloud coloring (see Figure 15 in Appendix C). Likewise, the word “feel” is grouped with all remaining terms, which includes the words with the lowest frequencies. As a result, I recommend using k-means clustering instead of hierarchical clustering, since the former method produces more effective groupings.

Conclusion

In this analysis, I incorporated several methods using SQL and R to perform a sentiment analysis on Claritin Twitter data within IBM’s DB2 Warehouse on Cloud service. My goal was to study patient sentiment levels and determine the impact of side effects towards user sentiment. Over the course of this analysis, I found that each of my methods provided useful information for understanding the sentiments of Claritin users. Firstly, my SQL queries were used to explore the dataset and find some of the relationships between variables. These queries revealed that side effects caused by Claritin were a rare phenomenon. However, if a patient were to experience any potential side effects, then their opinion towards Claritin would drastically decline.

My initial R methods involved creating a pie chart and two bar graphs. The pie chart was used to determine how often each side effect occurs. This image revealed that allergies are the most common symptom by far—taking up almost half of all side effect cases. Meanwhile, shortness of breath, nausea, heart palpitations, and headaches are the least common symptoms and consist of only a handful of cases each. My first bar graph displayed the distribution of side effects for each sentiment level. It revealed that allergies were by far the most common symptom amongst

users with the lowest sentiment levels—which suggests that allergies may produce the most severe symptoms related to Claritin usage. My second bar graph showed the relationship between a patient’s gender and the side effects they experienced. Based on this graph, female patients suffered far more than male patients from every recorded side effect. My findings suggest that women may be more vulnerable to the symptoms caused by taking Claritin.

I noticed two shortcomings regarding my implementation of these visualizations. First, these methods relied on the variable “symptoms,” which shows the name of the side effect (if any) that a patient has. I created this variable to allow all side effects to be included in the same graph, rather than creating separate graphs for each side effect. The issue with this method is that it only allows for one side effect to be associated with each user. If a patient has multiple symptoms, then one of the symptoms will overwrite the other. One recommendation would be to create a copy of any rows where the user has more than one symptom. This will allow for multiple side effects to be associated with the same user. The other limitation is that some side effects (such as headaches, shortness of breath, and nausea) only appear a handful of times in the dataset. Due to their small sample size, there may be insufficient information to determine how strongly they affect the user sentiment. For future analysis, I would recommend gathering a larger number of tweets to include more cases with these side effects. This may show us whether these symptoms have the same effect that “allergies” have towards user sentiment.

Next, I created a logistic regression model to predict whether patients will have a negative or non-negative sentiment. Based on the model results, I concluded that logistic regression is an effective method for predicting patient sentiment. This is because the model yielded a classification accuracy of approximately 88% in both the training and test data. Due to the training set and the test set having very similar results, there is no evidence of overfitting in this model.

Furthermore, the model sensitivity is between 85-95% in the training and test sets—which suggests that the model is very effective at identifying patients with negative sentiment. However, the main limitation is that the training sensitivity (93.6%) is noticeably higher than the test sensitivity (85.2%). The test results are still good, but there is room for improvement. One suggestion for further research is to remove insignificant inputs with p-values that are greater than 0.05. Removing unnecessary variables can improve the model performance (Knode, 2016), and thus I would recommend using this approach to produce a more accurate regression model.

From here, I incorporated text mining to evaluate the word usage in tweets with negative sentiment levels. The first text mining method involves creating a word cloud that shows the most frequent terms associated with negative sentiments. The most interesting finding from this method is that most words (such as “allergy,” “sinus,” and “headache”) relate to specific side effects. Based on this, it is apparent that side effect symptoms are the primary cause of low user sentiment. However, this word cloud also highlights an additional limitation in my approach. Since the method focuses solely on word counts, it offers little ability to understand the context of each word. For instance, I claimed that the word “work” may refer to whether a drug is working as expected. But since there is no context provided, we are only able to make educated guesses regarding each word’s usage and intended meaning. For future analysis, I would recommend using advanced NLP technologies such as IBM Watson. Watson’s cognitive computing approach can provide a deeper understanding of each word’s meaning, as well as its possible sentiment.

The final text mining component involved using both k-means clustering and hierarchical clustering to group terms into clusters based on their frequencies. My goal was to determine which of these clustering methods is a better fit for my dataset. Based on my findings, k-means clustering appears to fit the data more effectively than hierarchical clustering. This is because the k-means

clustering method grouped the words into appropriate groups that are consistent with my earlier findings on word frequency. Words with similar frequencies are grouped together, and the cluster with the most frequent terms by far has a greater distance from any other cluster. The hierarchical dendrogram, on the other hand, grouped terms with similar frequencies into separate clusters while placing too many terms into one massive cluster regardless of their frequency counts. As a result, k-means clustering is a more effective method for clustering the Claritin Twitter data.

Overall, the sentiment analysis results have yielded many insights for understanding why some users have negative opinions towards Claritin. Undoubtedly, side effects such as allergies are the main cause of low user sentiment levels. The patient's gender can also affect the likelihood of developing symptoms, and certain side effects may impact the sentiment more severely than others. This approach is not perfect, and my findings will likely improve with further research and the continued growth of sentiment analysis technology. But nevertheless, these strategies offer a good foundation for studying online user opinion towards Claritin and other medicines.

References

- Bati, F. (2015, Fall). Clustering. Lecture presented at UMUC. Retrieved July 9, 2017.
- Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). A Practical Guide to Sentiment Analysis (Vol. 5). Retrieved April 9, 2018, from <http://sentic.net/practical-guide-to-sentiment-analysis.pdf>
- Feinerer, I., & Hornik, K. (2017, March 2). Package 'tm'. Retrieved February 13, 2018, from <https://cran.r-project.org/web/packages/tm/tm.pdf>
- In Memory Analytics with IBM Db2 Warehouse on Cloud (2018). Retrieved April 9, 2018.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. Retrieved June 28, 2017.
- Knode, S. (2016, August 19). *Regression Models*. Lecture presented at UMUC. Retrieved October 3, 2017.
- Ng, A. (2017). *Lecture 81 - Choosing the Number of Clusters*. Lecture presented in Stanford University. Retrieved July 31, 2017, from <https://www.coursera.org/learn/machine-learning/lecture/Ks0E9/choosing-the-number-of-clusters>
- Oleson, D. (2013a, November 13). Claritin Twitter - dataset by crowdflower. Retrieved April 11, 2018, from <https://data.world/crowdflower/claritin-twitter>
- Oleson, D. (2013b, March 21). Discovering Drug Side Effects with Crowdsourcing. Retrieved April 11, 2018, from <https://www.figure-eight.com/discovering-drug-side-effects-with-crowdsourcing/>

Appendix A

Relevant IBM DB2 Warehouse on Cloud Images

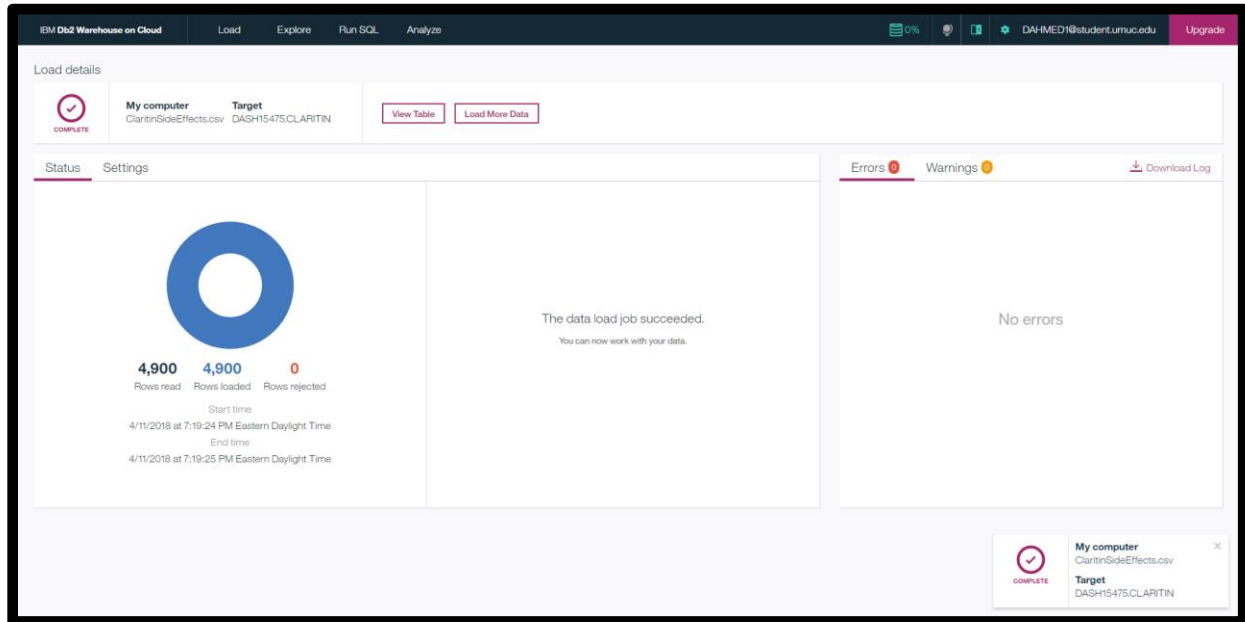


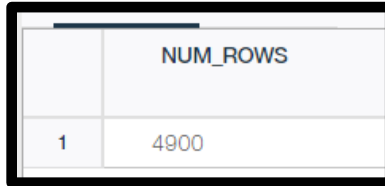
Figure 1. Data Load Status for Claritin Side Effects Dataset.

DASH15475.CLARITIN											Delete Table	Export to CSV
RELEVANT VARCHAR(1)	SENTIMENT SMALLINT	GENDER VARCHAR(10)	DIZZINESS VARCHAR(3)	CONVULSIONS VARCHAR(2)	HEART_PALP VARCHAR(3)	SHORT_BREATH VARCHAR(3)	HEADACHES VARCHAR(3)	DRUG_DECOR VARCHAR(3)	ALLERGIES VARCHAR(3)	BAD_INTER VARCHAR(3)		
yes	3	female	no	no	no	no	no	no	no	no		
yes	3	male	no	no	no	no	no	no	no	no		
yes	3	female	no	no	no	no	no	no	no	no		
yes	3	female	no	no	no	no	no	no	no	no		
yes	2	female	no	no	no	no	no	no	no	no		
yes	4	male	no	no	no	no	no	no	no	no		
yes	3	male	no	no	no	no	no	no	no	no		
yes	3	male	no	no	no	no	no	no	no	no		
yes	4	female	no	no	no	no	no	no	no	no		
yes	4	unknown	no	no	no	no	no	no	no	no		
yes	4	female	no	no	no	no	no	no	no	no		
yes	3	male	no	no	no	no	no	no	no	no		
yes	2	female	no	no	no	no	no	no	yes	no		
yes	4	male	no	no	no	no	no	no	no	no		
yes	3	female	no	no	no	no	no	no	no	no		
yes	4	female	no	no	no	no	no	no	no	no		
yes	3	female	no	no	no	no	no	no	no	no		
yes	3	female	no	no	no	no	no	no	no	no		
yes	4	bad_link_or_company	no	no	no	no	no	no	no	no		
non_english												
yes	4	female	no	no	no	no	no	no	no	no		

Figure 2. Claritin Side Effects Dataset.

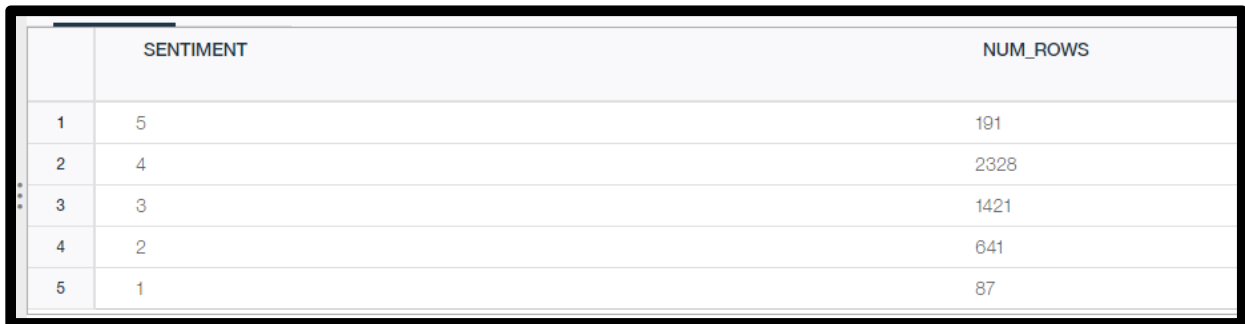
Appendix B

Relevant SQL Output Images



	NUM_ROWS
1	4900

Figure 1. Total Number of Rows in Claritin Dataset.



	SENTIMENT	NUM_ROWS
1	5	191
2	4	2328
3	3	1421
4	2	641
5	1	87

Figure 2. Number of Rows for Each Sentiment Level in Claritin Dataset.



	GENDER	NUM_ROWS	AVG_SENT
1	female	2557	3.368791
2	male	1558	3.493581

Figure 3. Number of Rows and Average Sentiment for Each Gender.

	SENTIMENT	NUM_ROWS
1	5	1
2	4	21
3	3	29
4	2	185
5	1	50

Figure 4. Sentiment Count for Patients Having any Possible Side Effect.

	NUM_ROWS	AVG_SENT
1	4382	3.49224

Figure 5. Count and Average Sentiment of Patients with no Side Effects.

	NUM_ROWS	AVG_SENT
1	286	2.083916

Figure 6. Count and Average Sentiment of Patients with at Least 1 Possible Side Effect.

	HEADACHES	SHORT_BREATH	ALLERGIES	NAUSEA	INSOMNIA	AVG_SENT
1	yes	no	no	no	no	2.265714
2	no	yes	no	no	no	2.25
3	no	no	yes	no	no	1.734848
4	no	no	no	yes	no	2
5	no	no	no	no	yes	2.384615

Figure 7. Average Sentiment of Patients with Specific Side Effects.

	ALLERGIES	SENTIMENT	NUM_ROWS	PERCENT
1	yes	4	2	1.515151
2	yes	3	3	2.272727
3	yes	2	85	64.393939
4	yes	1	42	31.818181

Figure 8. Sentiment Percentages for Patients with Worse Allergies After Taking Claritin.

	CONTENT	GENDER	SENTIMENT
1	@sawngbyrd28 Yes you can overdose on Claritin. Any signs of muscle contractions, drowsiness, or headache?	female	3
2	Went to the doctor...told him i had headaches and body aches...he gives me claritin -_-	female	2
3	Claritin just gave me migraine. Anybody else ever have that happen?	female	2
4	NOW i remember why i dont take claritin :(#massiveheadache	female	2
5	Just took some Claritin.. This headache is gettin on nerves!	male	3
6	Took a Claritin a few hours ago, yet my head still feels really stuffed up & in pain. "sigh"	male	2
7	Headache again, 2 hours into my day... looking into it, apparently claritin can do this. May have to choose between breathing and thinking	unknown	2

Figure 9. Tweet Content, Gender, and Sentiment Level for Patients with Headaches.

Appendix C

Relevant R Output Images

```
> # Checks the row counts in the new data frame and the database.
> nrow(CLAR_SE)
[1] 4900
> idadf(mycon, "SELECT COUNT(*) FROM CLARITIN")
      1
1 4900
```

Figure 1. Validation of Row Counts in SQL and R Data Files.

```
> # Lists all tables in the database.
> idaShowTables()
      Schema      Name      Owner Type
1 DASH15475 AIRLINE_SENTIMENT DASH15475 T
2 DASH15475 CLARITIN DASH15475 T
> # Verifies that the Claritin table exists.
> idaExistTable('CLARITIN')
[1] TRUE
```

Figure 2. Verification of Claritin Table's Existence.

```
> # Shows the number of tweets per sentiment.
> table(CLAR_SE$SENTIMENT)

 1    2    3    4    5
87  641 1421 2328 191
> idadf(mycon, "SELECT sentiment,
+ COUNT(1) AS count
+ FROM CLARITIN
+ WHERE sentiment IS NOT NULL
+ GROUP BY sentiment
+ ORDER BY sentiment DESC")
SENTIMENT COUNT
1          5  191
2          4 2328
3          3 1421
4          2  641
5          1   87
```

Figure 3. Validation of Number of Tweets per Sentiment Level.

```
> apply(CLAR_SE, 2, function(CLAR_SE) sum(is.na(CLAR_SE)))
```

CONTENT	TIME	RELEVANT	SENTIMENT	GENDER	DIZZINESS	CONVULSIONS	HEART_PALP	SHORT_BREATH	HEADACHES
0	0	0	232	232	232	232	232	232	232
HEADACHES	DRUG_DECR	ALLERGIES	BAD_INTER	NAUSEA	INSOMNIA				
232	232	232	232	232	232				

Figure 4. Number of Missing Values in Claritin Dataset.

	TIME	RELEVANT	SENTIMENT	GENDER	DIZZINESS	CONVULSIONS	HEART_PALP	SHORT_BREATH	HEADACHES
Wed, 24 Oct 2012 21:13:12	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Wed, 24 Oct 2012 23:58:54	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Thu, 25 Oct 2012 04:03:54	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Thu, 25 Oct 2012 04:06:52	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Thu, 25 Oct 2012 06:41:29	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Thu, 25 Oct 2012 12:08:59	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Thu, 25 Oct 2012 13:33:24	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Thu, 25 Oct 2012 14:42:22	+0000	no	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Tue, 16 Oct 2012 03:18:34	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Tue, 16 Oct 2012 16:39:11	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Tue, 02 Oct 2012 03:14:53	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Mon, 01 Oct 2012 03:50:36	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Mon, 01 Oct 2012 04:49:55	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Mon, 01 Oct 2012 04:56:19	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Mon, 01 Oct 2012 12:34:30	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Tue, 16 Oct 2012 22:49:52	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Tue, 16 Oct 2012 23:02:31	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
Sat, 20 Oct 2012 08:25:30	+0000	non_english	NA	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>

Figure 5. Preview of Rows with Missing Values in Claritin Dataset.

```
> summary(CLAR_SE)
```

CONTENT	TIME	RELEVANT	SENTIMENT	GENDER	DIZZINESS	CONVULSIONS
Length:4668	Length:4668	non_english: 5	1: 87	bad_link_or_company: 274	no :4657	no:4668
Class :character	Class :character	yes :4663	2: 641	female :2557	yes: 11	
Mode :character	Mode :character		3:1421	male :1558		
			4:2328	unknown : 279		
			5: 191			
HEART_PALP	SHORT_BREATH	HEADACHES	DRUG_DECR	ALLERGIES	BAD_INTER	NAUSEA
no :4663	no :4664	no :4661	no :4602	no :4536	no :4628	no :4664
yes: 5	yes: 4	yes: 7	yes: 66	yes: 132	yes: 40	yes: 26
						INSOMNIA
						no :4642

Figure 6. Descriptive Statistics of Variables in Claritin Dataset.

```
> summary(CLAR_SE$SYMPTOM)
```

allergies	bad interaction	dizziness	drug effect decreased	headaches
127	36	8	57	7
heart palpitations	insomnia	nausea	none	shortness of breath
4	26	4	4118	3

Figure 7. Number of Observations with Each Side Effect.

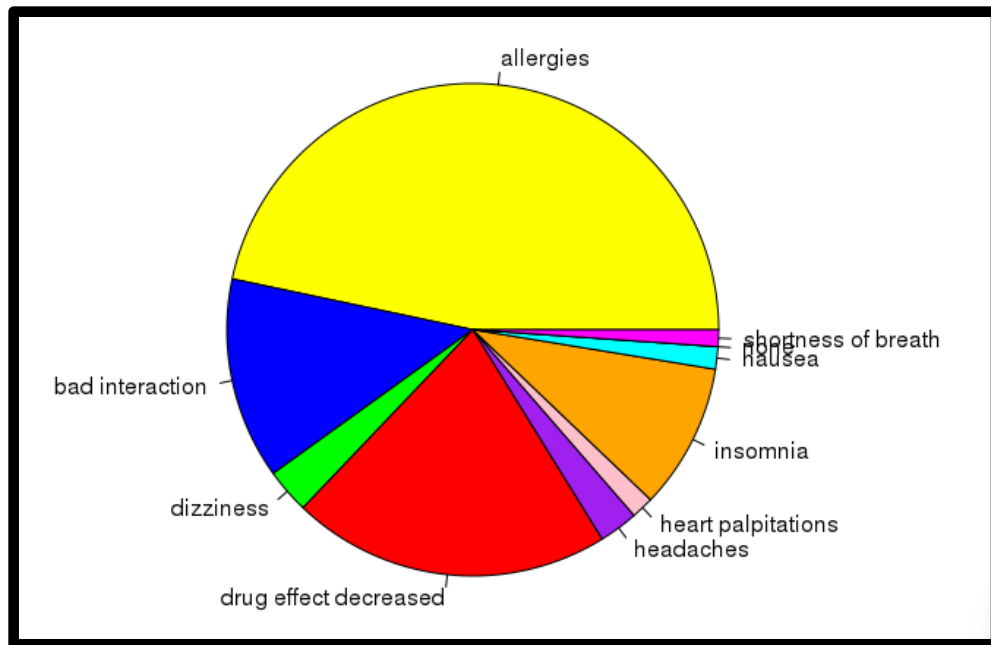


Figure 8. Overall Frequency of Each Side Effect.

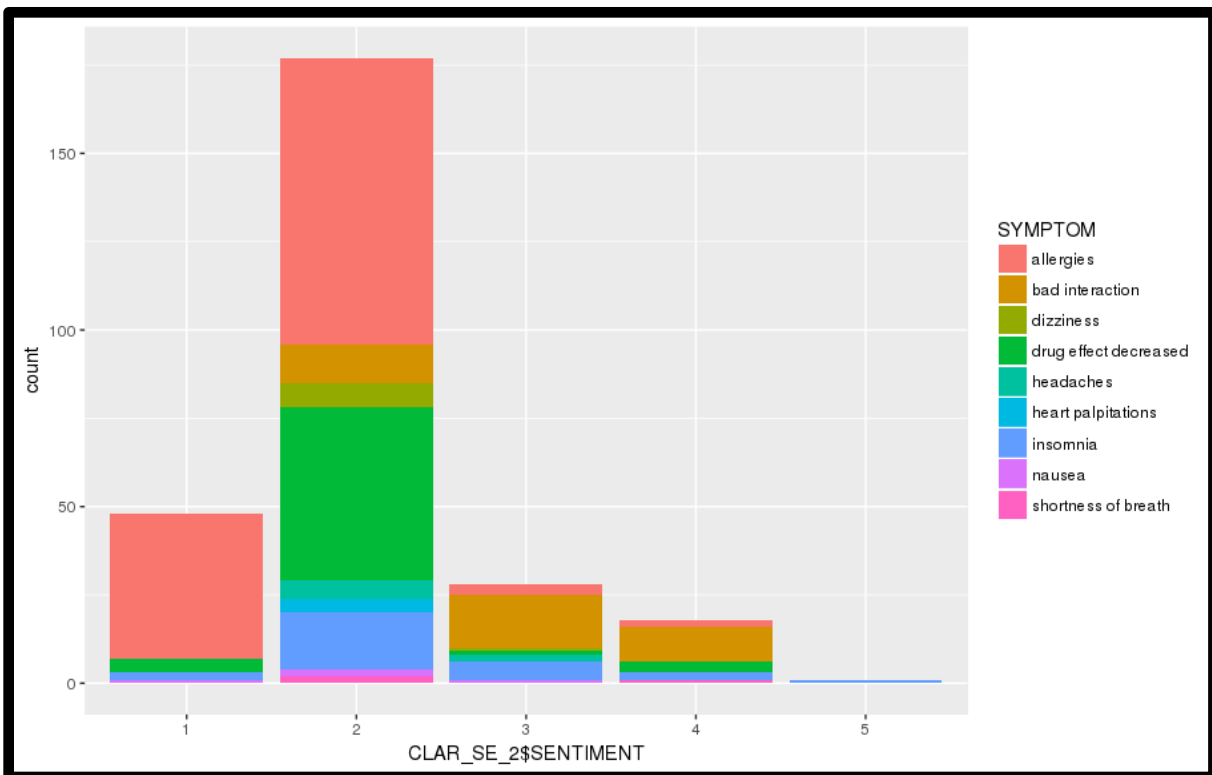


Figure 9. Frequency of Each Side Effect Based on Sentiment Level.

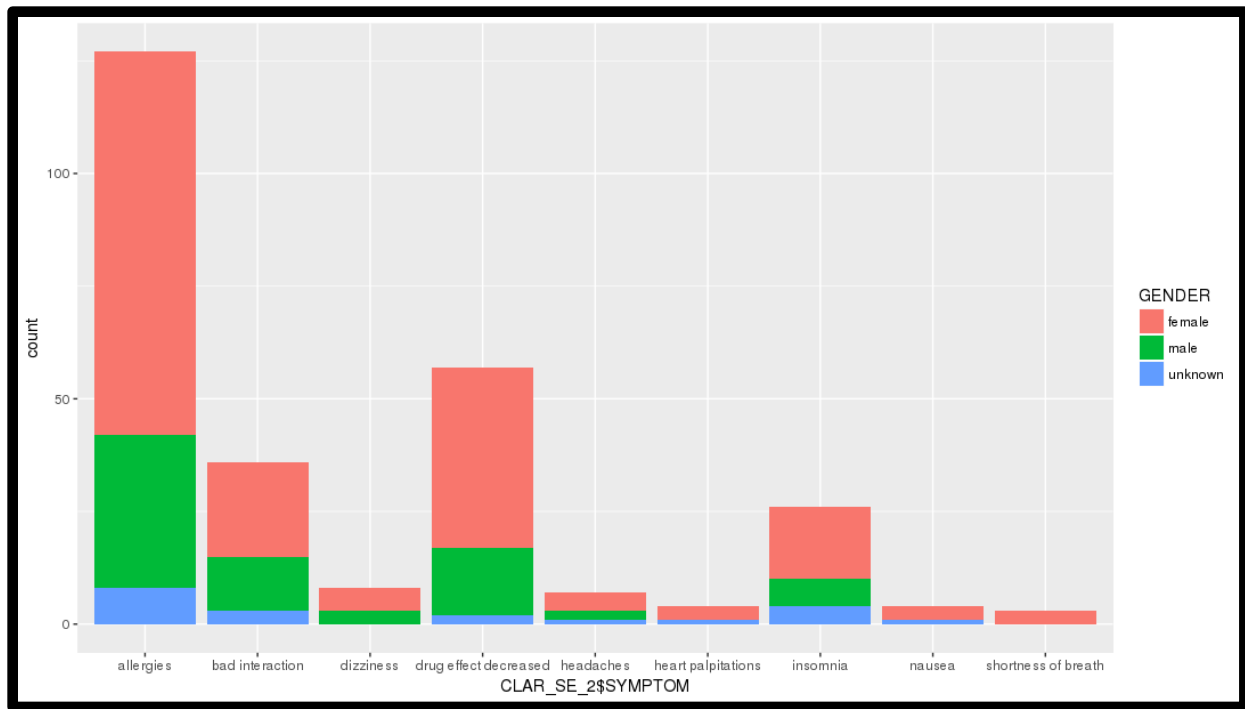


Figure 10. Number of Patients by Gender for Each Side Effect.

```
> table(train_pred, train.data$SENTIMENT)

train_pred    0    1
           0 2559  330
           1   10  146
```

Figure 11. Logistic Regression Confusion Matrix on Training Set.

```
> table(test_pred, test.data$SENTIMENT)

test_pred    0    1
           0 1112  152
           1   12   69
```

Figure 12. Logistic Regression Confusion Matrix on Test Set.


```

> # Shows the properties of the DTM after removing sparse terms.
> tweet_dtm_2
<<DocumentTermMatrix (documents: 728, terms: 39)>>
Non-/sparse entries: 990/27402
Sparsity           : 97%
Maximal term length: 9
Weighting          : term frequency (tf)

```

Figure 16. Properties of the Tweets DTM After Removing Sparse Terms.

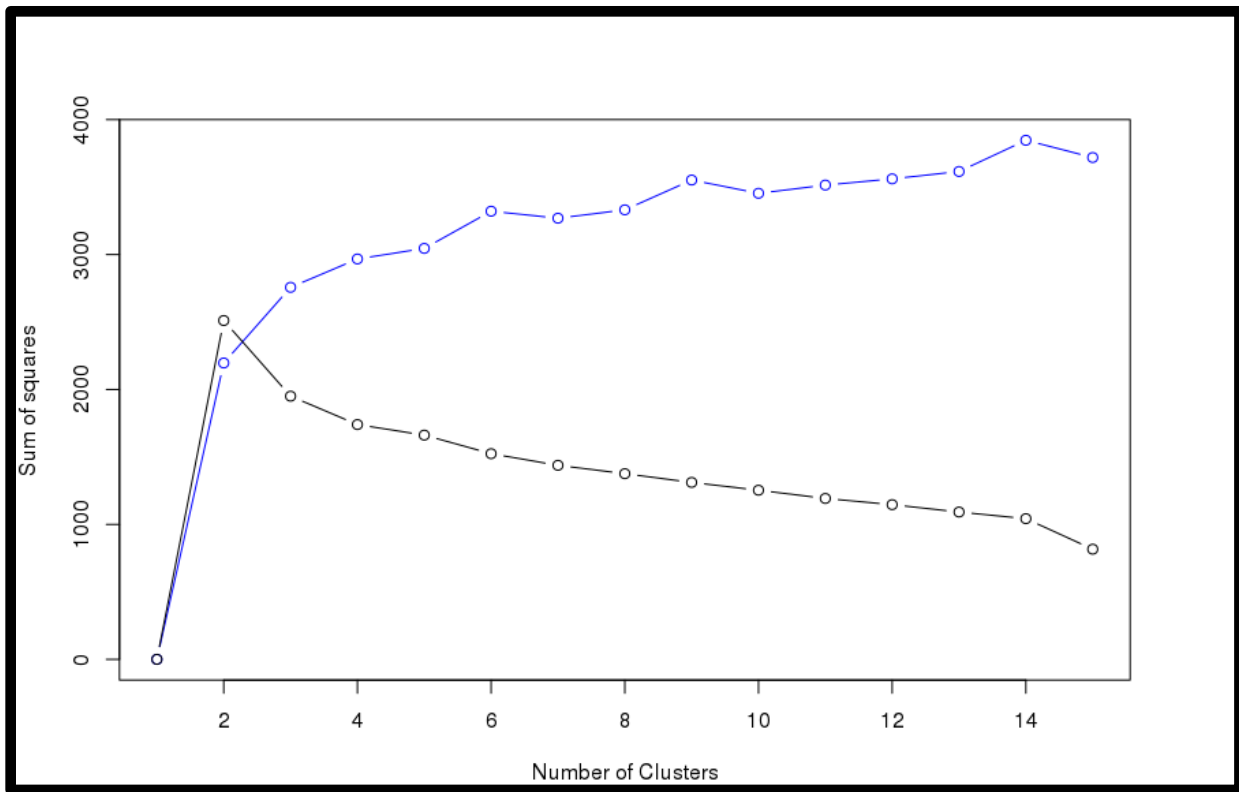


Figure 17. “Elbow Method” Visualization on Claritin Tweet Content.

```

> kfit
K-means clustering with 4 clusters of sizes 2, 4, 20, 13

Cluster means:
  hate  allegra  feel  mucinex  overdos  pill  allergi  hour  doe  work  hope  sick  dear
1 11.387457 11.443438 12.872618 11.610967 11.603278 12.158573 7.17635 12.606463 11.531829 7.17635 11.352350 11.778817 11.437171
2 8.233186 6.142475 7.164987 7.401678 7.768456 8.479040 12.03793 6.760274 7.912199 12.36046 7.501982 7.764928 7.433068
3 6.197237 7.029355 8.209755 4.829458 5.218185 6.723271 11.11369 7.267591 5.948401 12.06778 4.902122 5.229178 4.754059
4 6.552256 7.810532 8.907513 6.266290 6.590725 6.986477 11.51279 7.948013 6.480936 12.29410 6.288583 6.539883 6.115421

  day  medicin  nondrowsi  claritind  drug  sleep  tire  kick  sneez  benadryl  cough  eye  cold
1 11.855152 11.377251 11.609287 12.276794 11.564203 11.780579 11.698377 11.780579 11.991231 11.828459 11.612486 11.520797 11.564203
2 8.350485 7.568164 7.429598 8.404168 7.433497 7.633693 7.602929 7.858527 8.380051 8.384916 7.367293 7.536652 7.405655
3 6.481535 5.228254 4.853219 6.685826 5.008200 5.166195 4.997419 5.496656 6.495889 6.488433 4.790667 4.931483 4.853770
4 6.806882 6.564441 6.234199 6.945344 6.367038 6.436130 6.369512 6.778196 6.730449 6.795451 6.222616 6.159237 6.181980

  today  zyrtec  sinus  year  nose  ill  buy  drowsi  night  clear  morn  commerci  bad
1 11.741086 11.874267 12.031263 11.477226 11.872475 11.654009 11.826039 12.039016 11.911643 12.119863 11.655511 11.566066 11.345498
2 8.363821 6.386347 8.173464 7.499310 8.117964 7.428846 7.806628 8.079211 8.140025 8.354142 7.668625 7.303935 7.410600
3 6.535122 7.668912 6.180163 5.015604 6.217564 4.823275 5.417843 6.372227 6.048925 6.354555 5.110870 4.774008 4.920443
4 6.837582 8.371687 6.572511 6.368125 6.642964 6.208851 6.762318 6.775670 6.545959 6.689864 6.418913 6.126000 6.277154

Clustering vector:
  hate  allegra  feel  mucinex  overdos  pill  allergi  hour  doe  work  hope  sick  dear
  4      2      2      3      3      4      1      2      4      1      3      3      3
  day  medicin  nondrowsi  claritind  drug  sleep  tire  kick  sneez  benadryl  cough  eye  cold
  4      3      3      4      3      3      3      3      4      4      3      3      3
  today  zyrtec  sinus  year  nose  ill  buy  drowsi  night  clear  morn  commerci  bad
  4      2      4      3      4      3      3      4      4      4      3      3      3

Within cluster sum of squares by cluster:
[1] 220.7979 277.4257 574.5318 667.1523
(between_SS / total_SS = 63.0 %)

Available components:
[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size" "iter" "ifault"

```

Figure 18. Properties of the K-Means Clustering Model.

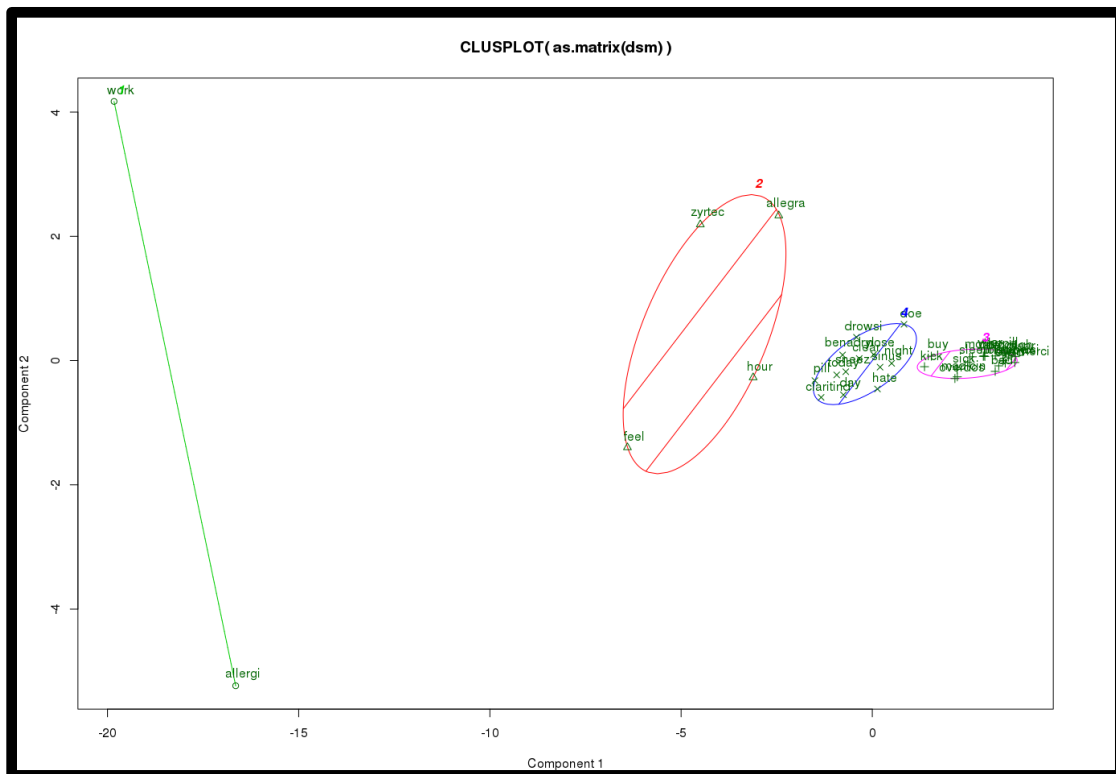


Figure 19. K-Means Clustering Model Using k=4 Clusters.

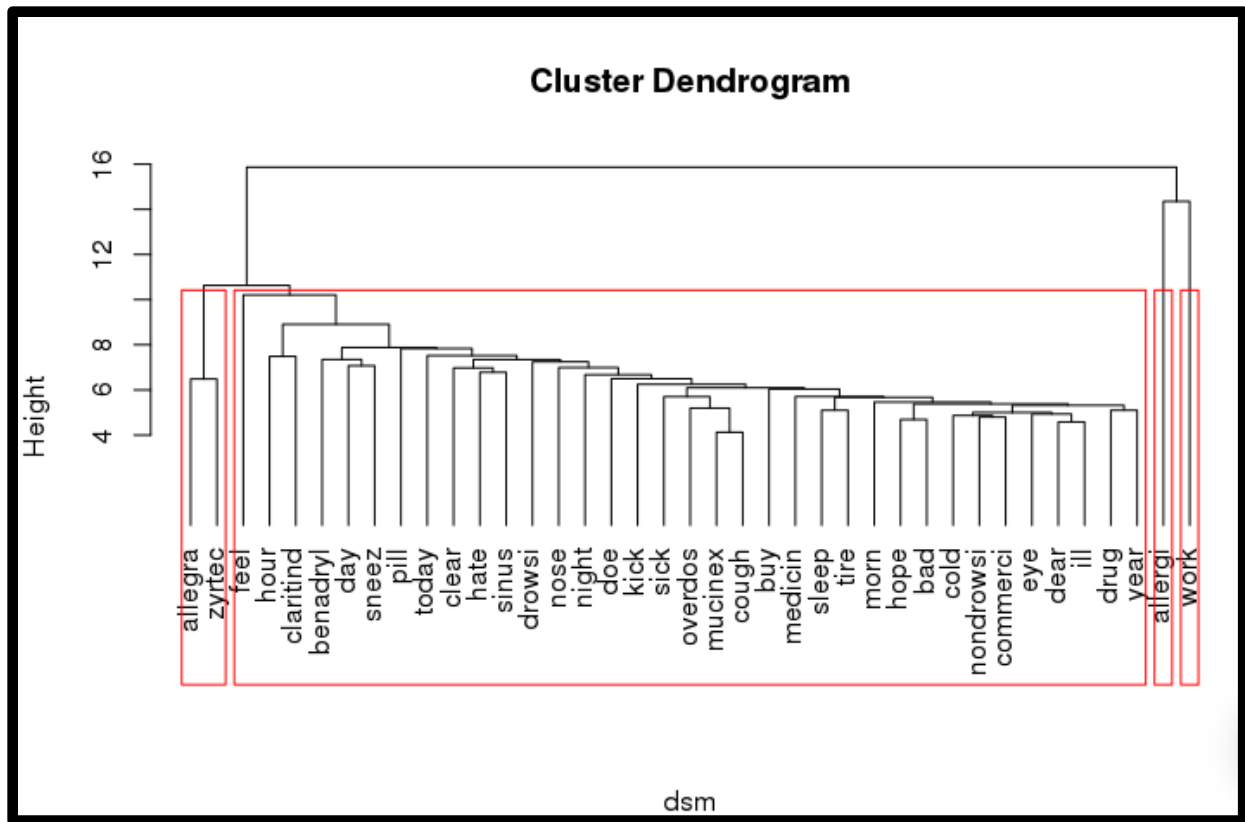


Figure 20. Hierarchical Clustering Dendrogram Using k=4 Clusters.