Final Data Exploration Report: Clustering Facebook Users Based on Statuses and Network Features to Analyze Big 5 Personality Traits

Kamryn Robertson and Laurel Smith

MATH 5330: Data Mining

Dr. Qiwei-Britt He

December 15, 2023

Introduction

In 2007, David Stillwell created the Facebook App myPersonality that allowed users the opportunity to take part in his psychological research by completing a personality questionnaire. With records of over six million participants, the myPersonality questionnaire rose to popularity on the Facebook application quickly. Users were provided scores and feedback on their personality type and even gave them the option to donate to the research, which in fact 40 percent gave back to the cause. This research project resulted in a mass interest in the topic of personality type and persuaded many other scientific journals to explore the topic and aid in the progression of retaining information on human psychology and behaviors. According to the myPersonality website, through their results Stillwell and Kosinki "provided other scholars with anonymized data they could use for non-commercial academic research which has allowed academics from many fields to make significant discoveries advancing our understanding of human behavior following the principles of open science and replicability". (Stillwell, D., & Kosinski, M., 2015) However, due to the amount of data collected and to ensure privacy, the data sharing with scholars was discontinued in 2018.

Dataset Background & Research Question

A workshop conducted in 2013 prepared the dataset used in this analysis that had the objective of setting a benchmark in personality prediction based on a user's Facebook statuses and social network features. It consists of the following: textual data (daily posts), individual's Big 5 personality dimensions, and social network (popularity among friends).

The Big 5 personality traits, which encompass the personality of an individual, are categorized into the following five groups:

1. Extraversion(x) (sociable vs shy)
2. Neuroticism(n) (neurotic vs calm)
3. Agreeableness(a) (friendly vs uncooperative)
4. Conscientiousness(c) (organized vs careless)
5. Openness(o) (insightful vs unimaginative)

Using the information of 250 Facebook users that took 100 item-long versions of IPIP questionnaires, 15 variables, and 9916 records of raw text data and network features (network size, betweenness, centrality, density, brokerage, and transitivity), categories (y/n) as the gold standard, and finally categorical variables in the Big 5 personality dimensions, we pose our research question as the following: How do Facebook statuses and network features contribute to the clustering of 250 users, and what insights can be gained by exploring the relationship between these clusters and the Big 5 personality traits?

Study Design

Our study design consisted of the following steps:

1. Analysis of Dataset
2. Data Preprocessing
3. Text Analysis
4. Selection and Analysis of Clustering Variables
5. Comparison of Big 5 Personalities with Clusters and Conclusions

Methodology

Using R 2023.06.2, we installed the following R packages to implement our code: ggplot2, tidyverse, tm, tokenizers, quanteda, topicmodels, slam, textTinyR, tidytext, syuzhet, dplyr, data.table, readxl, readr, ggcorrplot, cluster, dbscan, factoextra, "e1071", and ppclust.

Produced figures and results are presented in this paper, and supporting R code is supplied in a separate .Rmd file.

Data Analysis

The first step of our analysis was to study the features of the dataset. As stated in the introduction, our dataset had a total of 9916 records and this included the following variables:

1. #AUTHID: user's ID. We note that each user had multiple posts on multiple dates resulting in higher frequencies of a user's ID in some summaries.
2. STATUS: users' textual Facebook posts.
3. DATE: a date and time stamp of the Facebook post.
4. NETWORKSIZE: an indication of the number of direct connections between a user and its Facebook friends.
5. BETWEENNESS: an indication of the weighted number of shortest connected paths between a pair of Facebook friends.
6. NBETWEENNESS: a standardized measure of BETWEENNESS with a range of 0-100.
7. BROKERAGE: an indication of the number of connected Facebook friends a user does not have a direct connection with.
8. NBROKERAGE: a standardized measure of BROKERAGE with a range of 0-0.5.
9. Density: a calculation of the relationship between nodes in a network.
10. Transitivity: a percentage of a group of 3 connected Facebook friends in the network size.
11. cEXT: a categorical variable that indicates by yes or no if the user has an Extraverted(x) (sociable vs shy) personality. This variable is a gold standard.

12. cNEU: a categorical variable that indicates by yes or no if the user has a Neurotic(n) (neurotic vs calm) personality. This variable is a gold standard.
13. cAGR: a categorical variable that indicates by yes or no if the user has an Agreeable(a) (friendly vs uncooperative) personality. This variable is a gold standard.
14. cCON: a categorical variable that indicates by yes or no if the user has a Conscientious(c) (organized vs careless) personality. This variable is a gold standard.
15. cOPN: a categorical variable that indicates by yes or no if the user has an Openness(o) (insightful vs unimaginative) personality. This variable is a gold standard.

As our task was to find personality traits of users, and not individual statuses, we first had to modify our dataset to group multiple statuses together if they were associated with the same author. After doing this, we had a dataset with a dimension of 250 by 15. Once this was completed, to get a sense of the distribution of our gold standards, we plotted the overall personality traits among users, shown in **Figure 1**. Overall, users appear very insightful and calm, lean towards friendliness, are more shy than social, and are more careless than organized.

Data Cleaning

We removed columns 3 through 7, as these values were N/A and would not benefit our analysis. We also found that the dates were not read in as a date format. We fixed this by formatting all post dates to month/day/year. Overall, the data was nicely formatted and not much cleaning was required.

Text Analysis

We first added a column to our dataset. This column was a frequency column and calculated the number of total posts a given user created. We retrieved the frequency of each AUTHID or how many posts each user made from our dataset. By using *post_counts* we extracted from the first column how many times a user's ID was used for a post. We transformed this count into the data frame *post_counts_df.* The results of this are shown in the supporting R code, due to the size of the data frame.

Additionally, we retrieved the length of each post. To calculate the number of characters for each post we employ *nchar* on our dataset and also show these results in a dataframe. Similarly, the length of this result is substantially large and will be included in the supporting R code as well.

We then analyzed Unigrams and Bigrams to see the frequency and distribution of each single word and pair of adjacent words in our STATUS column. In both cases, we preprocessed the data by converting the letters to lowercase; removed punctuation, numbers, and stopwords; and stripped whitespace. For the unigram, we converted our text corpus to a document term matrix, and for the bigram, we converted our data to a document feature matrix.

The top n-grams found from our unigram are found in **Figure 2**. When we looked at the top results from our unigram/bigram combined data, we found that they were all still single words, as use of the top unigrams occur more frequently than the top bigrams.

Next, we conducted a sentiment analysis, relying upon three methods: syuzhet, bing, and afinn. According to the article "Text Mining and Sentiment Analysis: Analysis with R", the scale for sentiment scores using the syuzhet method is decimal and ranges from -1(most negative) to +1(most positive). (Mhatre, 2020). Also, we note that bing is a binary scale with -1 indicating negative and +1 indicating positive sentiment, and afinn is an integer scale ranging from -5 to

+5. (Mhatre, 2020) Utilizing the information from this article, we derived the following code to calculate our sentiment scores:

```
syuzhet_vector <- get_sentiment(data[i,2], method="syuzhet")
bing_vector <- get_sentiment(data[i,2], method="bing")
afinn_vector <- get_sentiment(data[i,2], method="afinn")
```

We added these calculated scores to our dataframe and normalized each column. These methods took care of data processing that we would normally consider in text mining, such as stopwords or stemming. In addition, there was a small amount of data cleaning involved in making sure the methods were given appropriate formats. In six observations, there were multibyte strings that the get_sentiment() function could not handle. We simply omitted these strings from the impacted rows to use get_sentiment(). One example of this is the following:

```
data[2,2] = "Supervisor: *PROPNAME* (second preference) Research Area: Regional Economic Integration
(fifth prefernece)" #removed multibyte in this row
```

Next we analyzed the correlation between the sentiments. As you would expect, these values were very highly correlated, as shown in **Figure 3** and **Figure 4**, where we showed the strength of sentiment correlation by using a “ggcorrplot” correlation matrix plot. The size and color intensity of the squares in the plot correspond to the magnitude of the correlations. We validated these results by the following code:

```
correlation = cor(sentiment)
print(correlation)
##                 syuzhet_sentiment bing_sentiment afinn_sentiment
## syuzhet_sentiment         1.0000000      0.7752712       0.8077621
## bing_sentiment            0.7752712      1.0000000       0.7421381
## afinn_sentiment           0.8077621      0.7421381       1.0000000
```

Here, we see that the syuzhet and afinn sentiment are the most correlated, so we choose afinn as our sentiment analysis method.

Prior to performing our clustering analyses, we need to ensure that we are using good variables. By looking at the pairs plot and constructing a correlation matrix (shown in **Figure 5** and **Figure 6**), we see that several variables are highly correlated. So, we remove some of the variables that have a high correlation with another variable in our dataset. As shown in **Figure 7**, this leaves us with NETWORKSIZE, NBROKERAGE, TRANSITIVITY, Freq, and afinn_sentiment, which we collated and named cluster_final.

Cluster Analysis: K-Means

For our clustering analyses, we had to choose an optimal k number of clusters. To accomplish this, we plotted the within-cluster sum of squares (WCSS) for k values between 2 and 10. Our results are shown in **Figure 8**. Based on this, the elbow is clearly at k=4, so we chose that value as our optimal k.

For our first clustering method, we utilized K-Means.

```
km = kmeans(cluster_final, centers = 4, nstart = 5)
```

We got four clusters of sizes 101, 15, 90, and 44. Then, we accessed the WCSS and between-cluster sum of squares (BCSS) from the results of our previous k-means clustering code. We achieved the following results respectively:

```
## [1] 660899.3 437427.9 826153.9 690461.2
## [1] 22823404
```

The second cluster has the lowest SSE, but it is also a lot smaller than the other clusters.
We also evaluated the quality of the K-Means clustering by the silhouette width of each cluster, shown in **Figure 9**. The average silhouette width is 0.5, which we concluded to be reasonable. We also looked at the pairs plot by clusters, shown in **Figure 10**, and made an observation that the nextwork size strongly dictated which cluster was which, when compared with the other four columns in cluster_final. Finally, we plotted our final K-Means cluster results by their first two

principal components, shown by **Figure 11**. This was not strictly necessary, as we only had five attributes, but it provided another visualization for us as we sought to understand our findings.

Cluster Analysis: Fuzzy C-Means

Next we used the Fuzzy C-Means clustering algorithm. We use k=4 as before.

```
cluster_cmeans <- cmeans(as.matrix(cluster_final), centers = 4)
```

This returned four clusters of sizes 15, 88, 51, and 96. This is rather similar to our K-means clustering results, with one of our cluster sizes being identical (15). As before, we evaluated the quality of our clustering by the silhouette width, shown in **Figure 12**. Once again, we got a silhouette width of 0.5. Then, we plotted the pairs plots (**Figure 13**), noticing the same behavior as before, and finally plotted the clusters using the first two principal components (**Figure 14**). All C-Means figures are very similar to their K-means counterparts. This almost identical behavior was an indication to us that we selected an optimal k for our clusters and that the clusters were grouped appropriately.

Results

After clustering our data by the K-Means and Fuzzy C-Means methods, we were able to see the distribution of the Big Five personalities for the users in each cluster. Since our clusters were so similar, the bar plot results (**Figure 15** and **Figure 16**) look almost identical. For the purposes of explaining the results, we will refer to the clusters from the K-Means method, shown in **Figure 15**.

The following table shows a summary of each cluster and the percent “yes” that each cluster displays according to the specified personality.

K-Means Cluster Number	**cEXT (%)**	**cNEU (%)**	**cAGR (%)**	**cCON (%)**	**cOPN (%)**

1 (Top Left)	19.8	53.5	45.5	44.6	71.3
2 (Top Right)	66.7	26.7	66.7	66.7	80
3 (Bottom Left)	41.1	32.2	62.2	52.2	70
4 (Bottom Right)	65.9	27.3	50	63.6	65.9

Cluster 1, with 101 users, is most defined by its shyness. However, this cluster is also very insightful. Users in this group are much more neurotic than in other clusters, while being less agreeable and less conscientious. Overall, this group can be characterized as shy, neurotic, uncooperative, and careless, while also being insightful.

Cluster 2, with 15 members, is our smallest group. Users of this cluster are very calm and are the most open, extroverted, agreeable, and organized group. Their non-neurotic tendencies likely allow them to be organized and might play into the high levels of insightfulness that this group scores. Further, their high levels of extraversion and agreeableness pair well together. As a whole, this cluster is our extroverted, open, friendly, and calm group. The size of this cluster might impact why it was able to achieve such a polarizing proportion of either “yes” or “no” tallies.

Cluster 3, composed of 90 members, does not hold the highest or lowest score for any one category. They are generally open, calm, shy, and agreeable. In addition, they tend to be rather organized.

Finally, Cluster 4, with 44 users, is rather similar to Cluster 2. Users split the vote on agreeableness, but other than that, we can consider this cluster to be calm, extroverted, open and organized.

Discussion

The most striking part of our study is the similarity between the two clustering algorithms. Upon analyzing both clusters, we saw that our clusters were very similar, with our principal components cluster plots looking identical to each other and our silhouette plots reflecting very similar clusters. One of the clusters was the exact same for both the K-Means and the Fuzzy C-Means clustering, and the other three clusters only varied by a few observations between the methods. This indicated that we could be confident in our clustering results.

When we analyzed the clusters, we found that all four clusters scored high on openness. This made it less helpful to distinguish between clusters for openness, although the high percent of "yes" values in Cluster 2 made us consider this the most open group. Further, users were generally agreeable and organized, which makes sense, as they chose to participate in this personality study.

Our study does have some limitations. First, participants had to fill out a questionnaire, meaning that our gold standards are likely self-selected and therefore biased. Second, participants had to agree to be a part of the study, meaning that results will be skewed towards people willing to participate in studies. In addition, we only relied upon two clustering methods. Though both gave similar results, the differences in cluster sizes could impact the validity of our data, as 15 is a rather small cluster size. Future work could use an additional clustering algorithm to support or inform our results. However, through this analysis, we were able to characterize how different clusters scored on the Big 5 personality traits.

Conclusion

By using the provided data, composed of raw text, network features, and questionnaire results, as well as performing sentiment analysis, we were able to cluster 250 Facebook users

into four groups using the K-Means and Fuzzy C-Means clustering algorithms. We found that the most important features involved NETWORKSIZE, NBROKERAGE, TRANSITIVITY, Freq, and afinn_sentiment. Then, we saw that NETWORKSIZE was able to clearly separate the clusters when paired with the remaining features. Finally, we compared our clustering results to the Big 5 personality traits and drew conclusions about the personalities of users in each cluster.

References

Stillwell, D., & Kosinski, M. (n.d.). MYPERSONALITY PROJECT. MyPersonality.org. Retrieved December 12, 2023, from https://sites.google.com/michalkosinski.com/mypersonality

Mhatre, S. (2020, May 13). Text Mining and Sentiment Analysis: Analysis with R. Redgate. Retrieved December 12, 2023, from https://www.red-gate.com/simple-talk/databases/sql-server/bi-sql-server/text-mining-and-sentiment-analysis-with-r/

Appendix

Figure 1. Distribution of Yes/No Values Across Personality Traits Across all Users

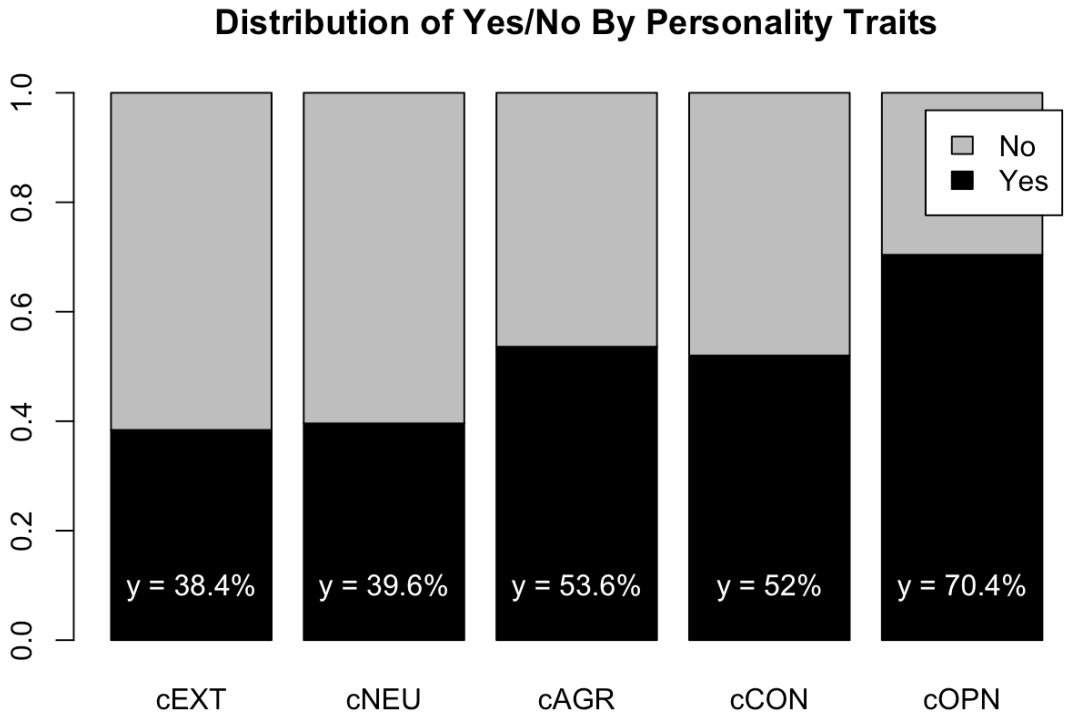

Figure 2. Top N Unigrams

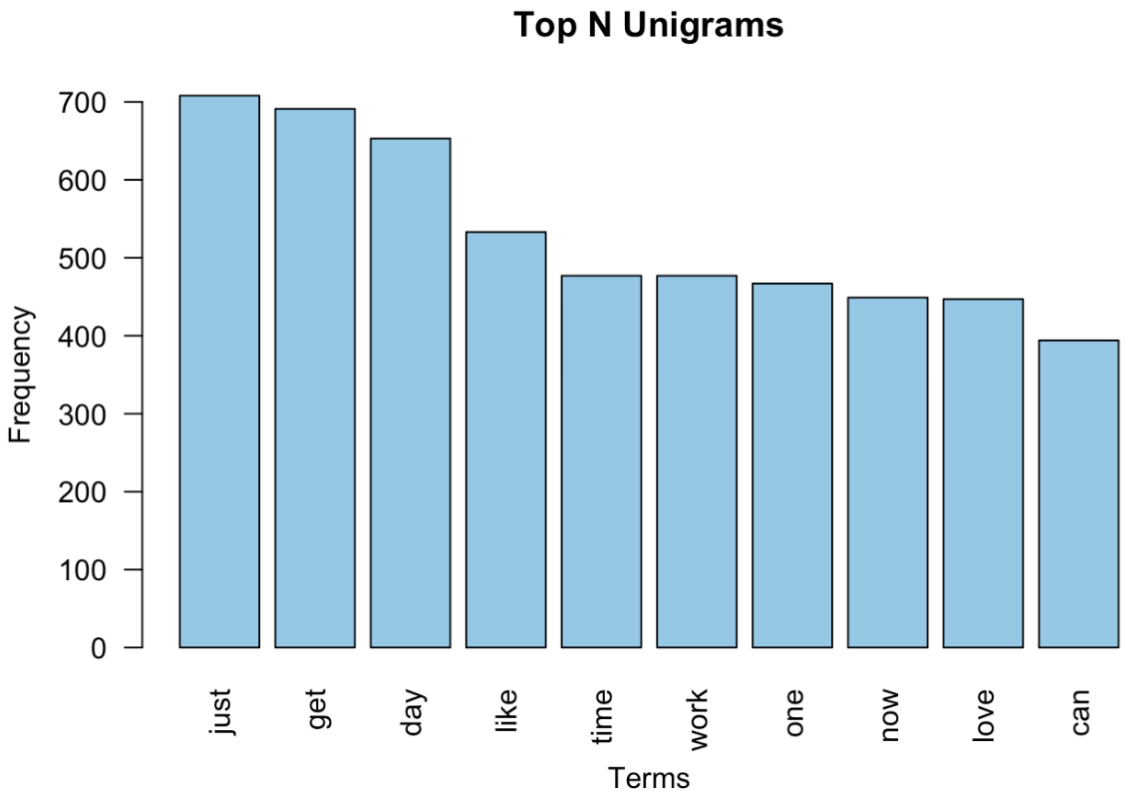

Figure 3. Pairs of Sentiment Scores

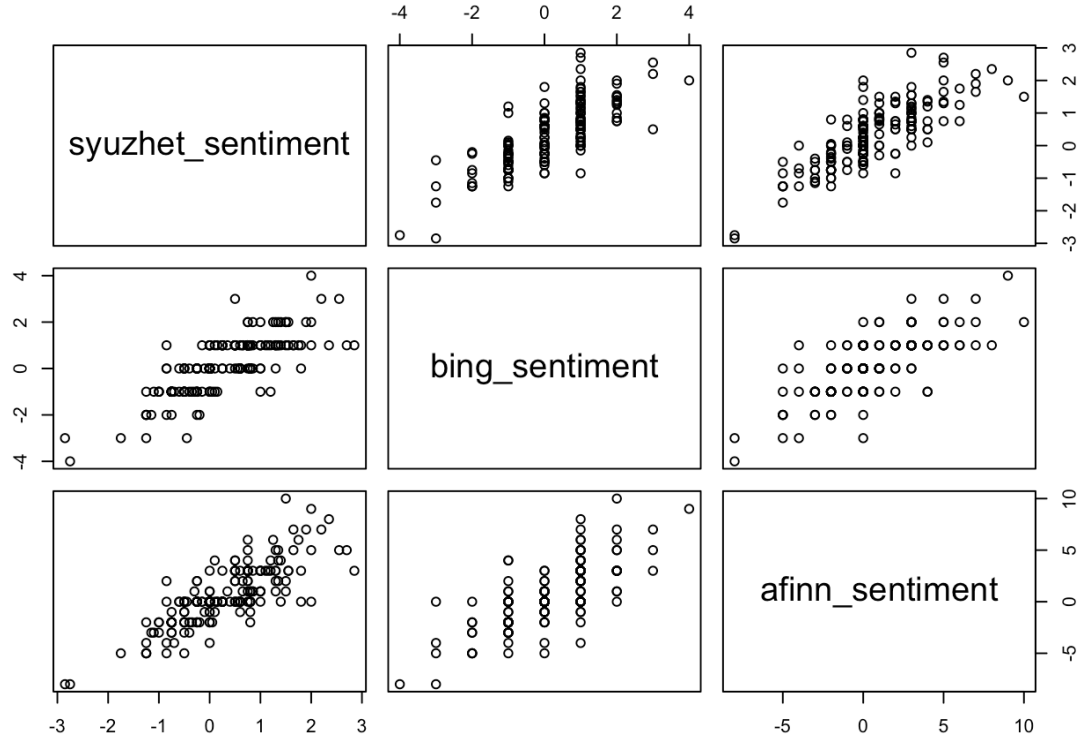

Figure 4. Correlation Between Different Sentiment Methods

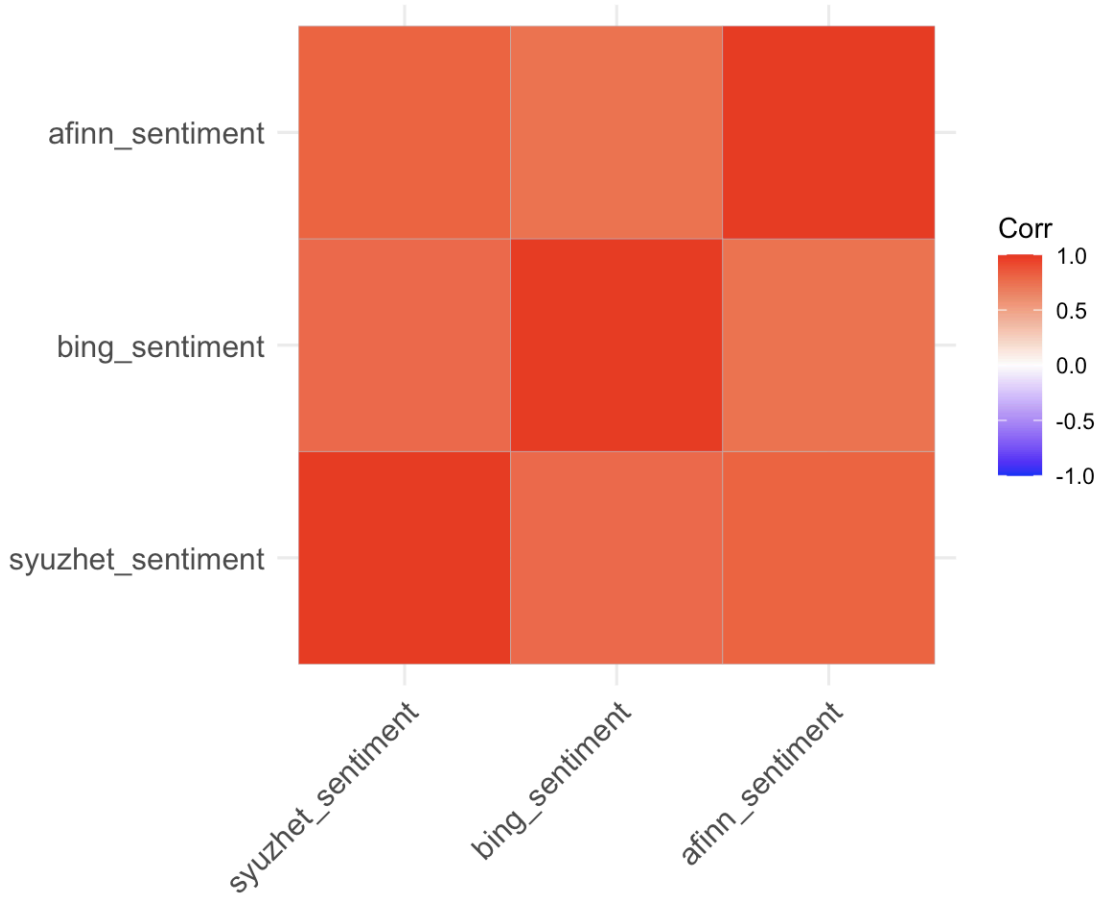

Figure 5. Pairs of Initial Clustered Data Features and Sentiment Scores

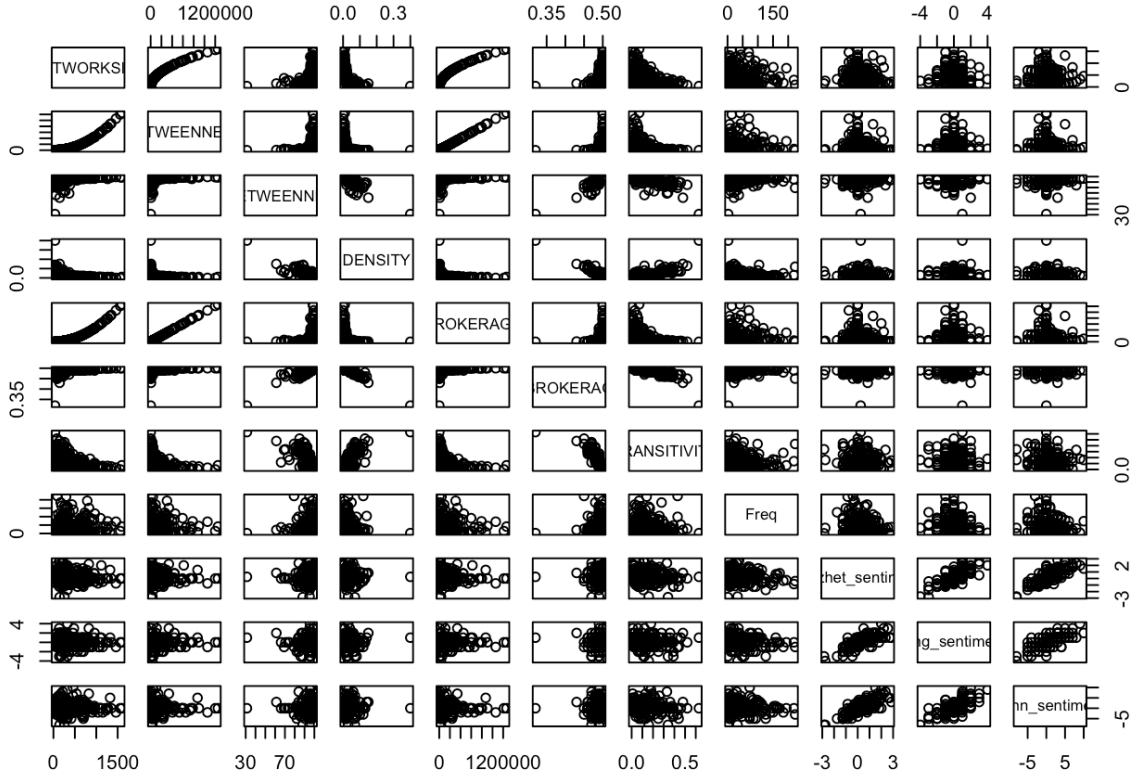

Figure 6. Initial Cluster Correlation

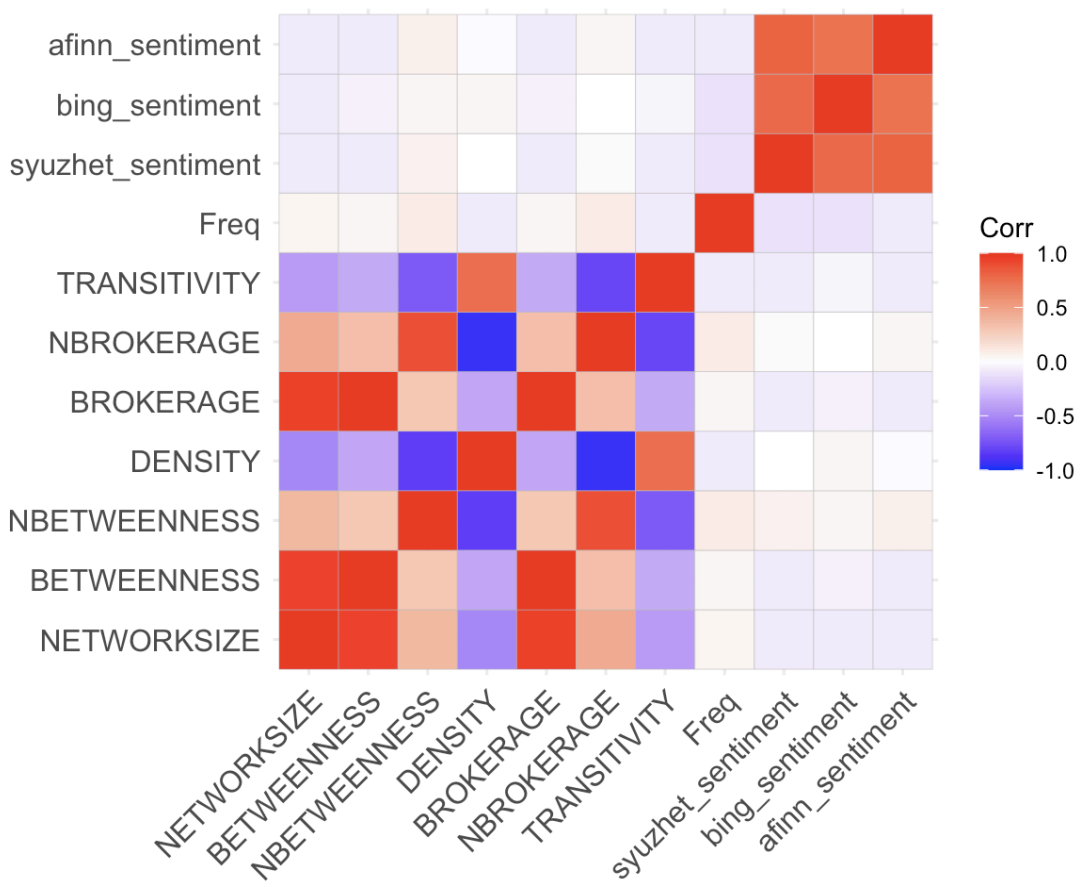

Figure 7. Final Cluster Correlation

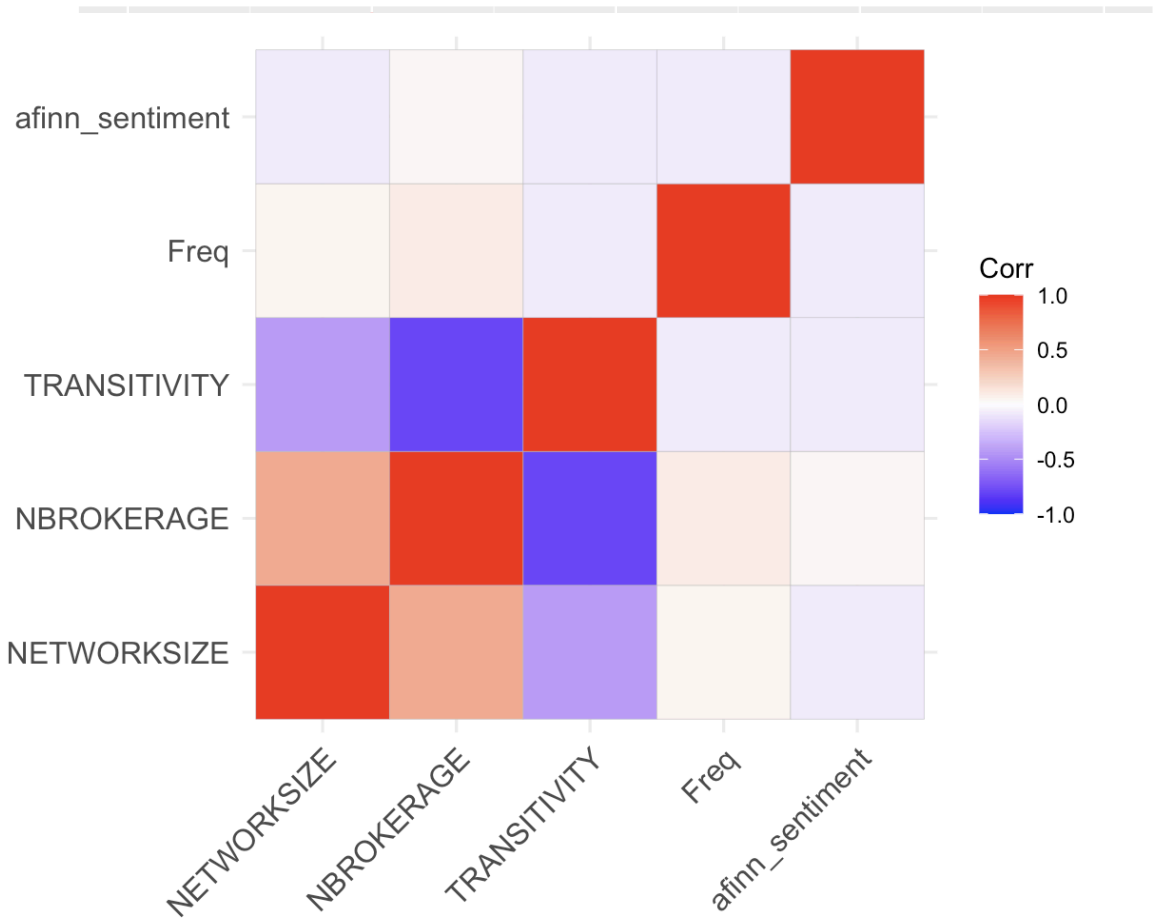

Figure 8. K-Means WCSS Plot

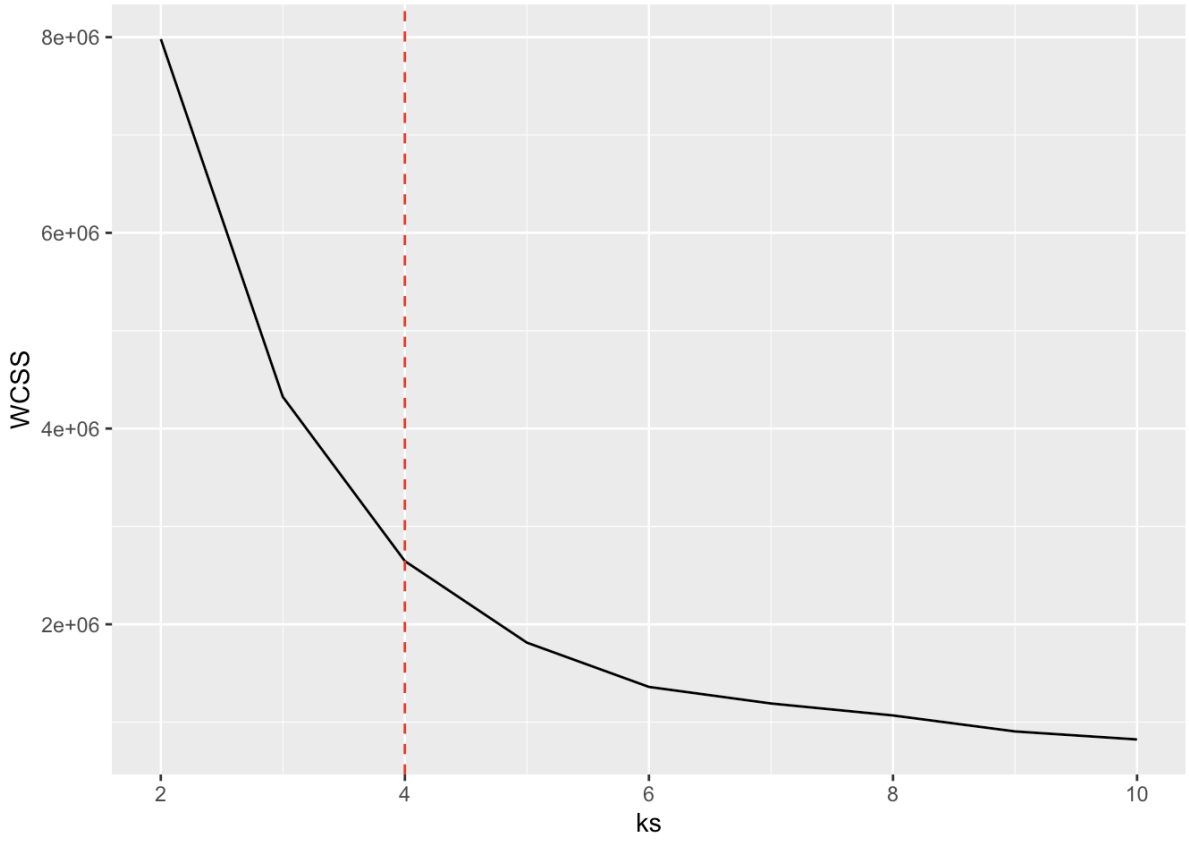

Figure 9. K-Means Silhouette

Figure 10. K-Means Pairs Plot

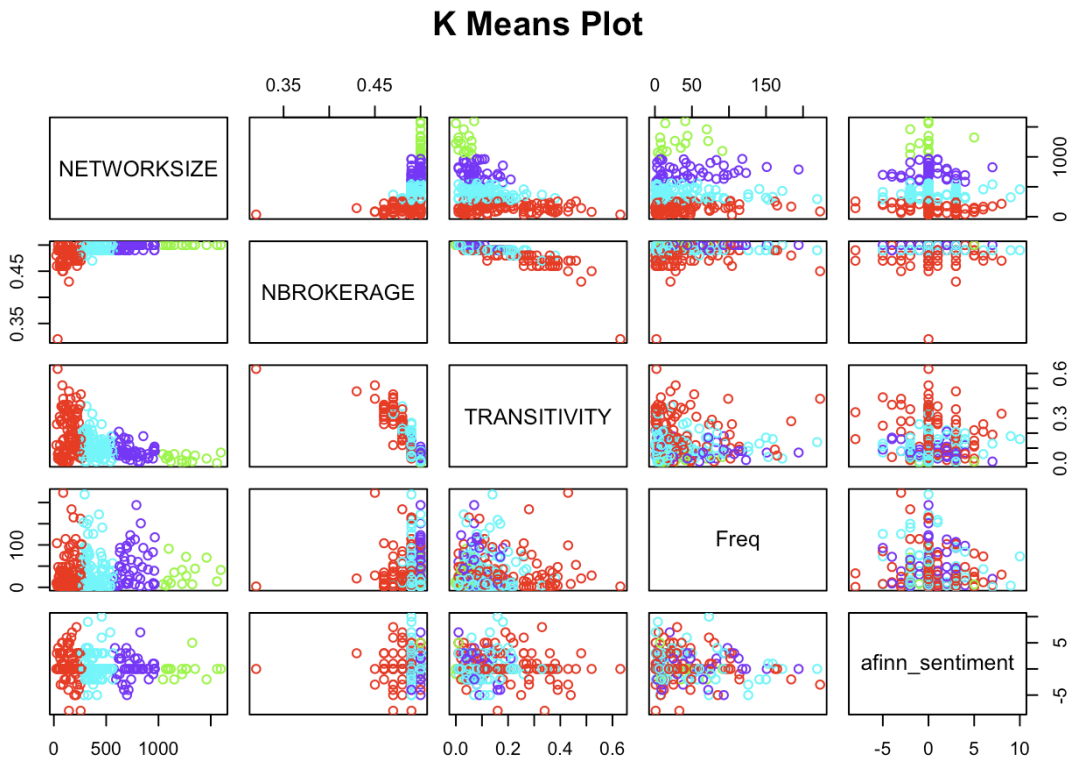

Figure 11. K-Means Cluster Plot

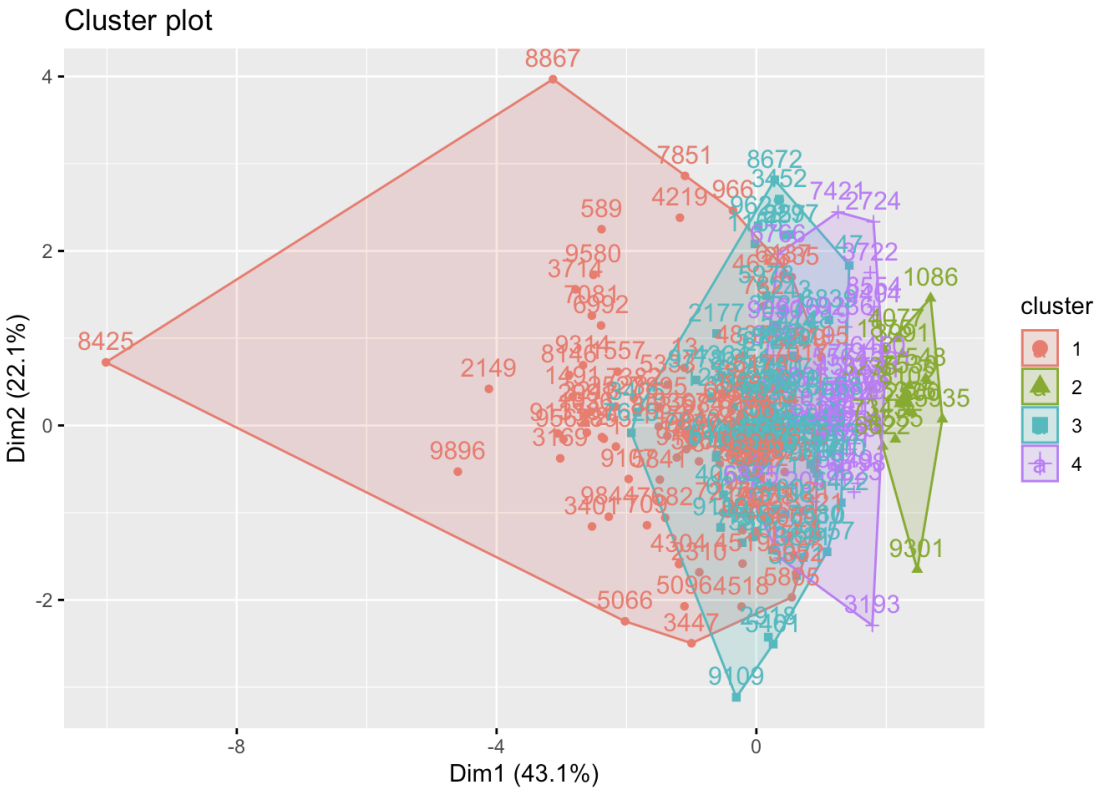

Figure 12. Fuzzy C-Means Silhouette

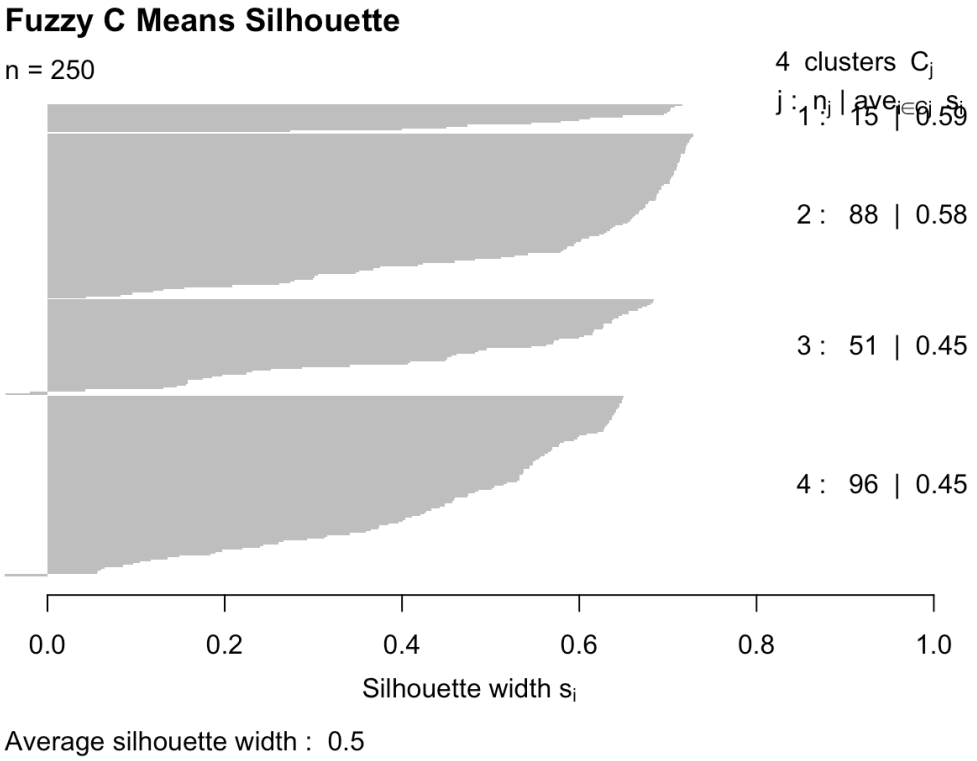

Figure 13. Fuzzy C-Means Pairs Plot

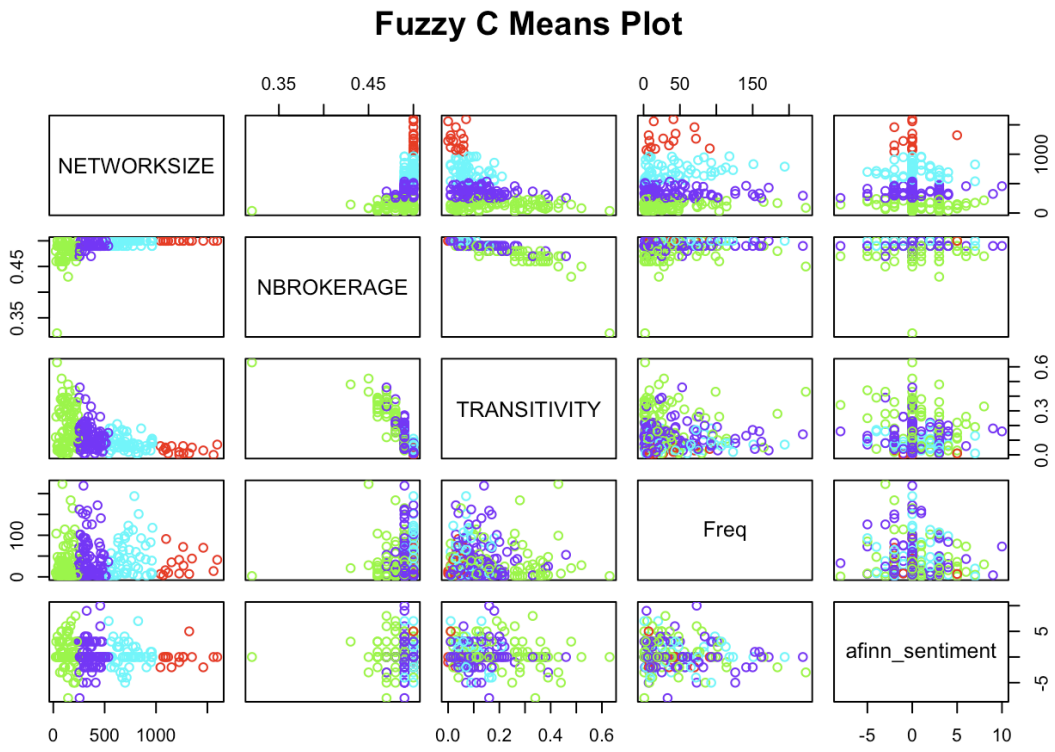

Figure 14. C-Means Cluster Plot

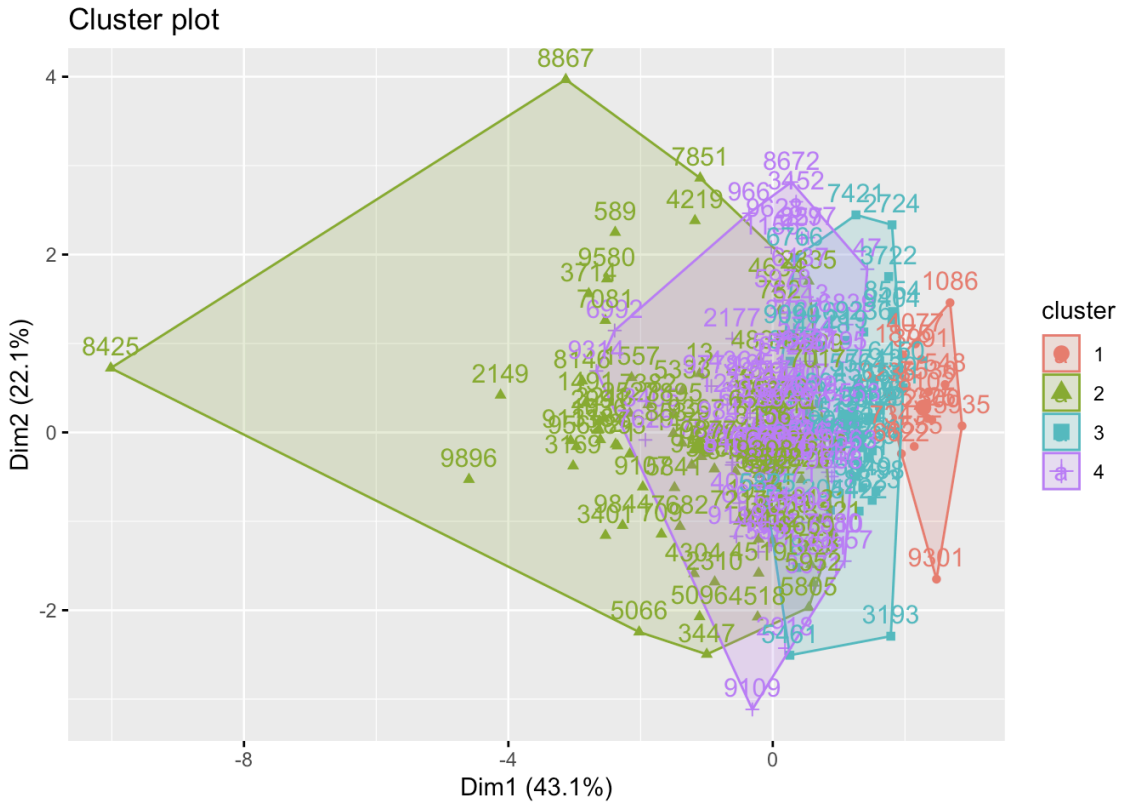

Figure 15. Distribution of Big Five Personalities for each K-Means Cluster

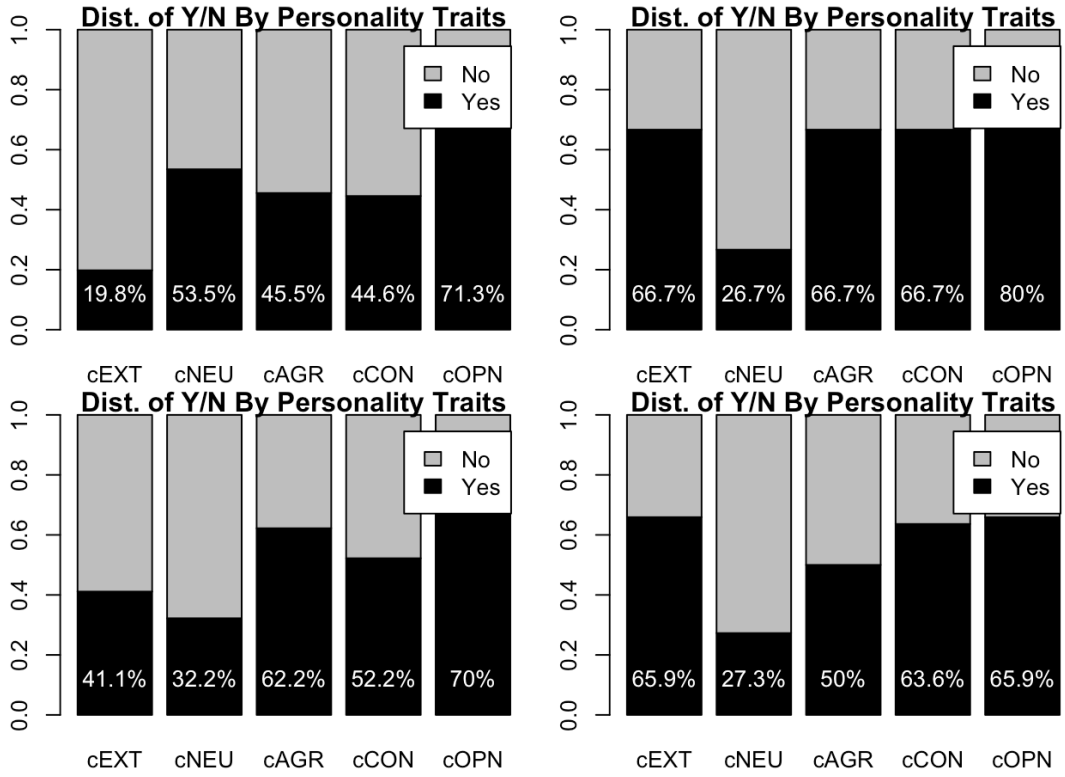

Figure 16. Distribution of Big Five Personalities for each Fuzzy C-Means Cluster