Math 656 Data Mining Final Project

Cansu Freeman and Cecilia Ma

December 2022

1 Background

The myPersonality dataset was obtained by the Psychometrics Center at the University of Cambridge during 2008-2012. The Center collected Facebook status and personality data from Facebook users in that time frame. In our project, we use a subset of this data, comprised of 250 users' statuses and personality traits based on the five-factor personality model, also known as the Big5. The Big5 is the most widely used and well-researched model for personality to date, and consists of five main traits: Neuroticism (neurotic vs calm), Extraversion (sociable vs shy), Openness (insightful vs unimaginative), Agreeableness (friendly vs uncooperative), and Conscientiousness (organized vs careless), (Chmielewski et al, 2013). Each individual scores a certain level for each of the five traits, which when combined will paint a larger picture of an individual's personality. Our project seeks to use textual data from user's Facebook statuses to gain an understanding of user's personality.

2 Data Source(s)

2.1 Data Description

The dataset provided consisted of 15 variables and 9917 records for 250 users. Each record was an individual Facebook status, containing censored user identification, the date, the

individual's network size, several other network details, along with the five personality traits. The traits are listed as either yes or no rather than a numerical scale. For example, if a user had sociable tendencies, the user's extroversion score would simply be "yes." In the dataset, there was one single missing value out of 9916, so we removed that row. The user had several other posts. From here, we combined the dataset by each user's identification code. We also changed the identification code to be a simple numerical integer for ease of use throughout the project.

2.2 New Variables

We created several new variables per user: total and average words, total and average posts, total and average ego words, and conducted sentiment analysis. The ego words are "me-words" that refer to when the user writes about his or herself in the status. Because this data was collected in the very early days of Facebook, we noticed that users often referred to themselves in the third person. For example, one user made the status, "runs like a girl, but she runs fast," referring to herself. Thus our "me-words" comprised of the following: 'me', 'my', 'myself', 'mine', 'we', 'our', 'ours', 'ourselves', 'his', 'hers'. For each user, we counted the total number of times the me-words, then averaged the me-words per number of posts.

For the sentiment analysis variable, we used the Transformers library with Python, which provides a sentiment-analysis API for classifying text sequences to positive or negative sentiments. We used this library to evaluate each individual status to determine if it was positive or negative (note: not neutral). The output result for each comment was a label of either positive or negative, and a numerical score representing the strength of the sentiment. In the interest of time and because the algorithm used a lot of computing power, we separated only the label data and gave it a binary score of 0 and 1, where 1 is considered a positive sentiment. We then averaged the sentiment score per user.

In the end, we have a dataset comprised of 250 rows, and 8 numerical columns: network size

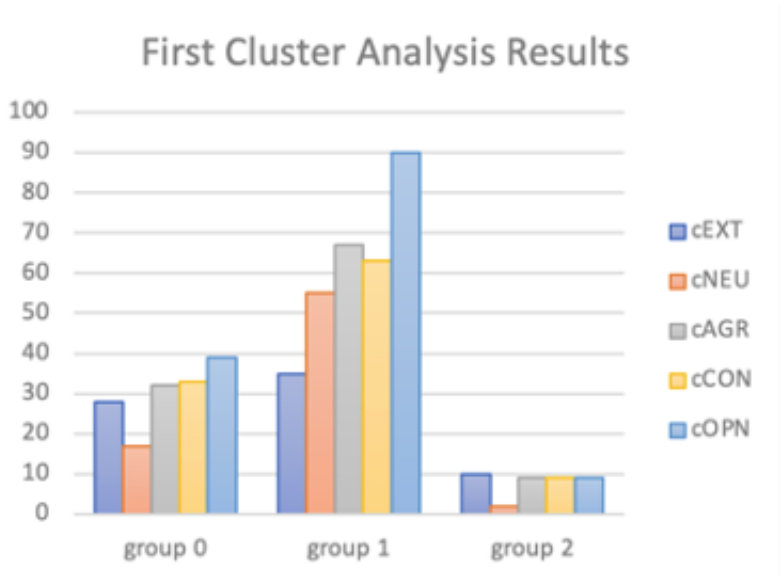
scaled, n-betweenness, density, n-brokerage, transitivity, average words, me-words per post, and sentiment. Now we will proceed with modeling.

3 METHODS AND ANALYSIS

3.1 K-means Clustering

Initially, we will try K-means clustering method on the entire dataset with all 8 variables. When selecting the appropriate number of clusters, we used the elbow method. We tested from 1 to 11 clusters, and the elbow method at this stage determined the "knee" to be at 3.

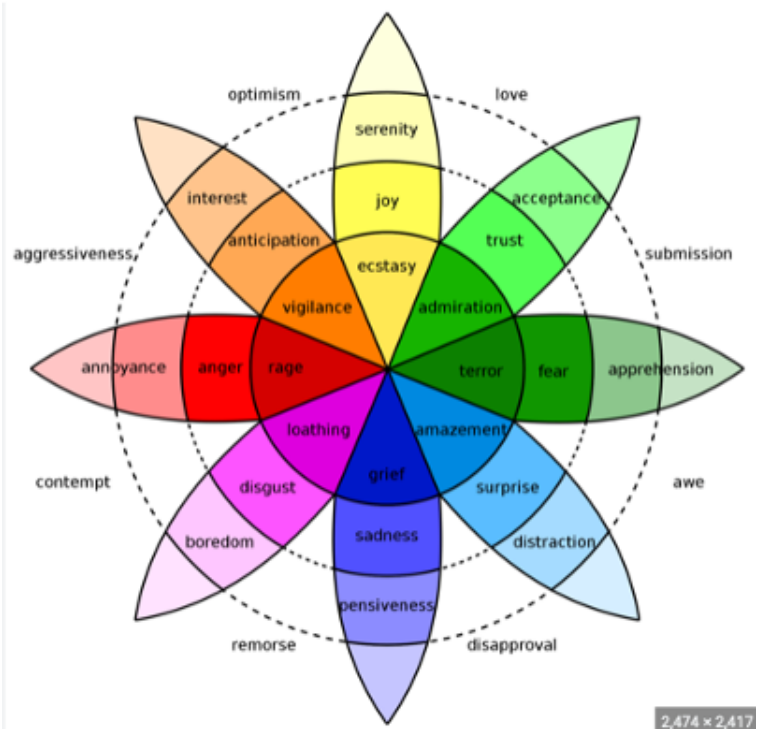
After clustering into three groups, each of the groups had the following composites of the different personalities. Unfortunately, because every group had similar proportions of the five personalities, it was hard to discern any real differences. The one possible inference is that the cluster named "group 2" had very little neurotic users, but group 2 was small, containing only 14 users. Next, we will add text mining analysis and then conduct principle component method before trying to cluster again.



3.2 Feature Extraction

3.2.1 NRC Emotion Lexicon Introduction

NRC Emotion Lexicon lists associations of words with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive), and it is created by manual annotation on a crowd-sourcing platform. Here is a picture of how NRC Emotion Lexicon divide sentiments and emotions:



Also, in NRC lexicon, one word could have both positive and negative sentiment associations, and could also have more than one emotion associations. The words 'abundance' and 'absence shown below give example about one words with different emotion and sentiment:

Word-Sentiment Associations	
abscess	negative
absence	negative
absent	negative
absentee	negative
absenteeism	negative
absolute	positive
absolution	positive
absorbed	positive
absurd	negative
absurdity	negative
abundance	positive negative
abundant	positive

Word-Emotion Associations	
abscess	sadness
absence	fear sadness
absent	sadness

3.2.2 Text Cleaning

We combined each user's text together, and clean it before extracting features. We converted all strings to lowercase, replaced contraction to their multi-word forms, replaced informal writing with known semantic replacements, replaced internet slang to normal words, removed hashtag, number, URL, HTML and punctuation. Also, we changed all emoticon marks to words. Finally, all removed stopwords were removed for all text.

3.2.3 Emotion Pairing

After cleaning the text, we paired the NRC Emotion Lexicon sentiment 'negative', 'positive' and emotion 'anger', 'anticipation', 'disgust', 'fear', 'joy', 'sadness', 'surprise', 'trust' with the tokenized and cleaned text. Also, as we mentioned above, each words can have multiple 'sentiment' labels. Here is a sample of paired words:

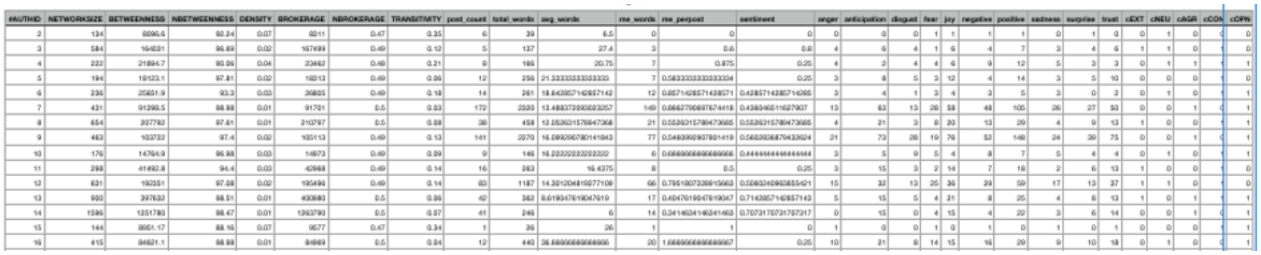
token <chr>	sentiment <chr>
sore	anger
sore	negative
sore	sadness
base	trust
hurting	anger
hurting	fear
hurting	negative
hurting	sadness
nun	negative
nun	trust

We created 10 variables corresponding these emotions and sentiment, and count the sentiment and emotion frequencies for each user individually. However, there are 11 users whom their texts are meaningfulness, which, for example, their text is full of URL and stop-words, so we cannot pair their texts with NRC emotion Lexicon. Therefore, we decided to

remove these 11 users in the dataset.

ID <chr>	anger <dbl>	anticipation <dbl>	disgust <dbl>	fear <dbl>	joy <dbl>	negative <dbl>	positive <dbl>	sadness <dbl>	surprise <dbl>	trust <dbl>
10	3	5	9	5	4	8	7	5	4	4
100	5	4	1	1	10	10	9	2	4	4
101	0	2	0	0	1	1	1	1	1	1
103	8	28	9	11	39	26	51	17	13	22
104	8	15	6	8	13	27	26	10	7	10
105	0	0	0	0	1	0	2	0	0	1
106	0	0	0	0	0	0	1	0	0	0
107	0	4	1	0	5	0	5	0	4	5
108	5	37	1	7	27	17	56	9	13	21
109	1	1	1	1	2	2	4	1	0	5

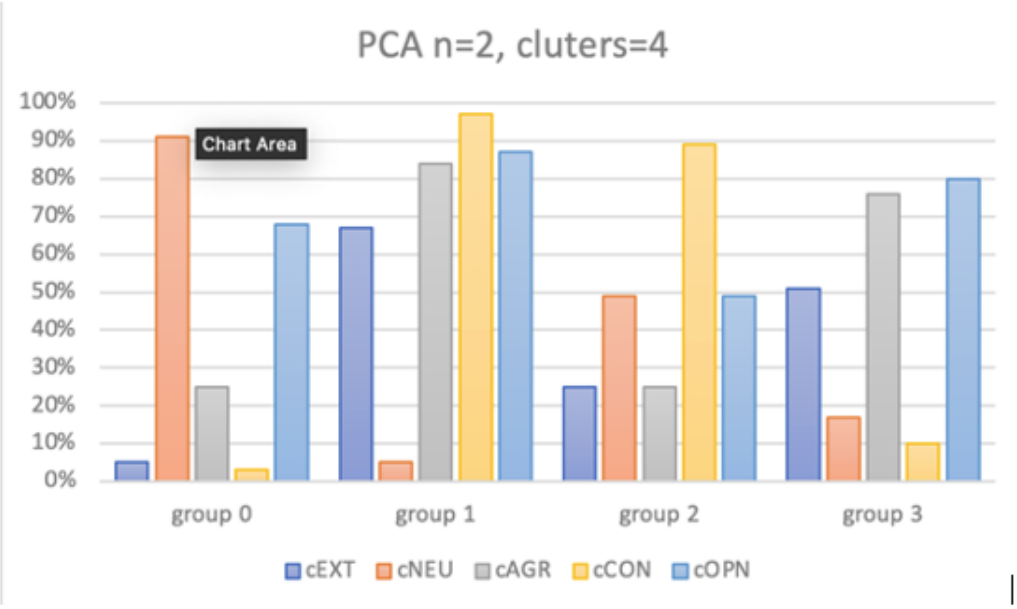
Now, The dataset has 23 attributes (except ID and 5 personalities' attributes) and 239 observations as shown below:



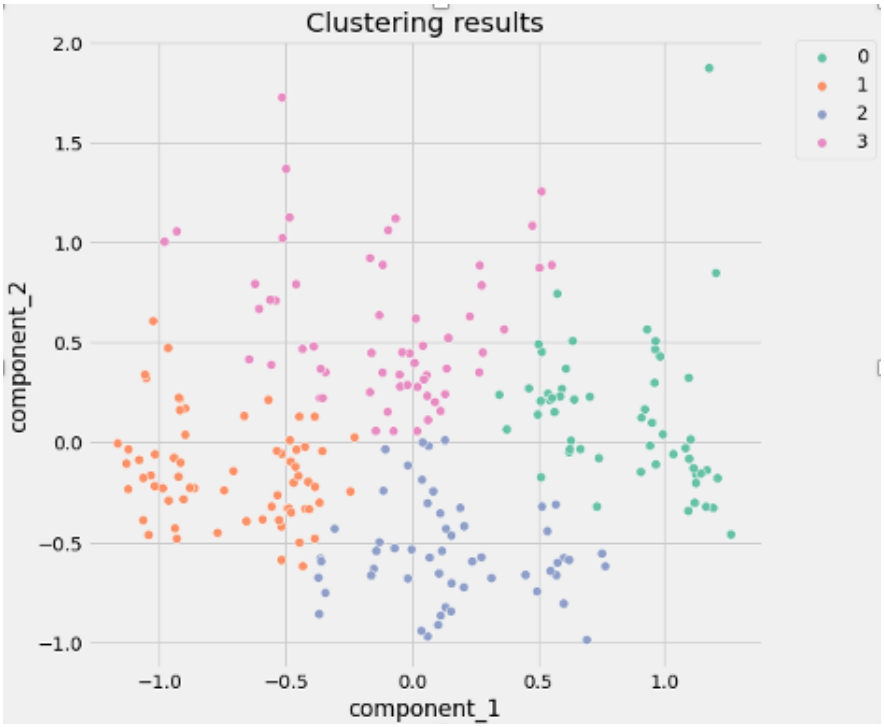
3.3 K-means Clustering cont.

With new features obtained from text mining, we will now attempt to conduct a K-Means clustering algorithm again with principal component analysis. PCA reduces the data down to two dimensions. Here we see clearer distinctions between some of the clusters. For example, in group 0, 91% of the users have neurotic as a trait, only 5% are extroverted, and 3% have conscientiousness. People with high neuroticism and low conscientiousness are usually regarded as having health-risk behaviors such as smoking and overeating, (Sutin et al, 2010). Likewise, people with high neuroticism and low extraversion may be "at higher risk for falls and mobility decline," (LeMonda et al, 2015). This finding suggests that there

is some association between the overall sentiment and feelings of one's online posts to health conditions.



In contrast, in group 1 only 5% of the users have neurotic trait, while 97% have conscientiousness. High conscientiousness combined with low neuroticism is a recipe for healthy lifestyle habits. Conscientiousness refers to one's self-discipline, abilities to make plans and stick to them, organizational skills, and cautiousness. Group 1 represents individuals who are well rounded, able to handle conflict, and are open to new experiences. When looking at a graph of the clustering results, we can see that group 0 is on the far right and group 1 is on the far left. This suggests that there is somewhat of a pattern or scale across the two components, where the left side has individuals who are more likely to be healthy minded, and the right side has individuals who are more likely to have poorer mental health outcomes.



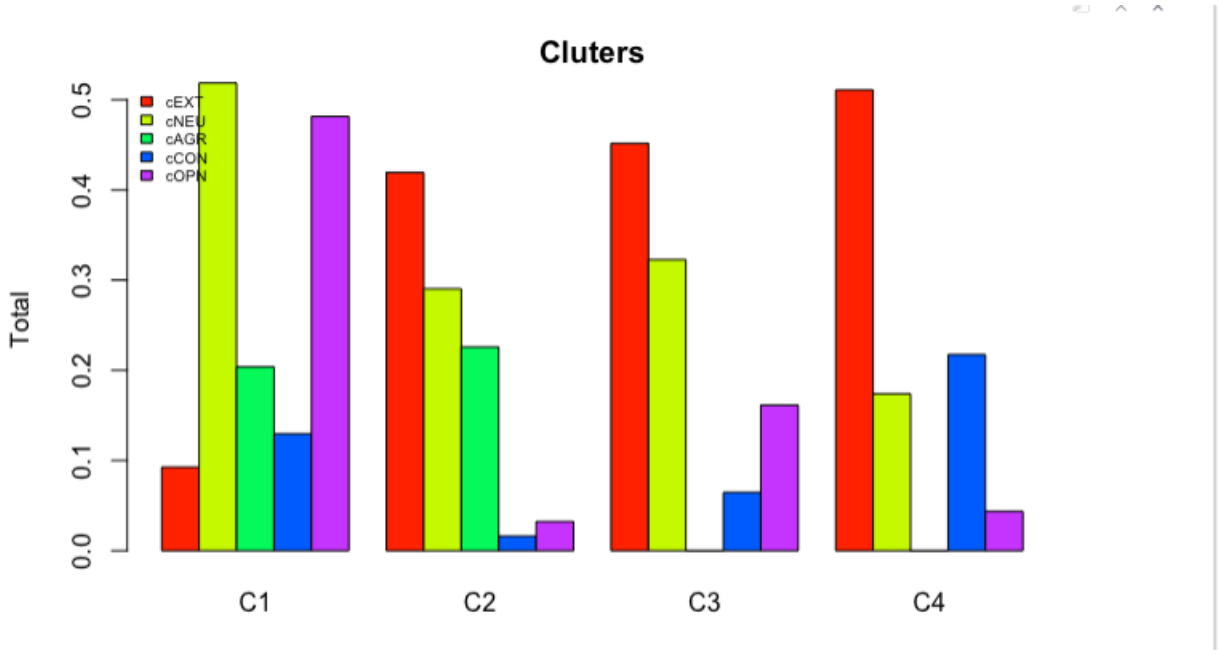
3.4 Fuzzy C-Means Clustering

We also tried the fuzzy C-Means clustering for 4 clusters with all variables, and the size for each cluster was mostly balanced at 54, 62, 31, 92. The (between_SS / total_SS) equals 0.44. Here is snippet of the data frame that shows the personalities' proportion in each cluster.

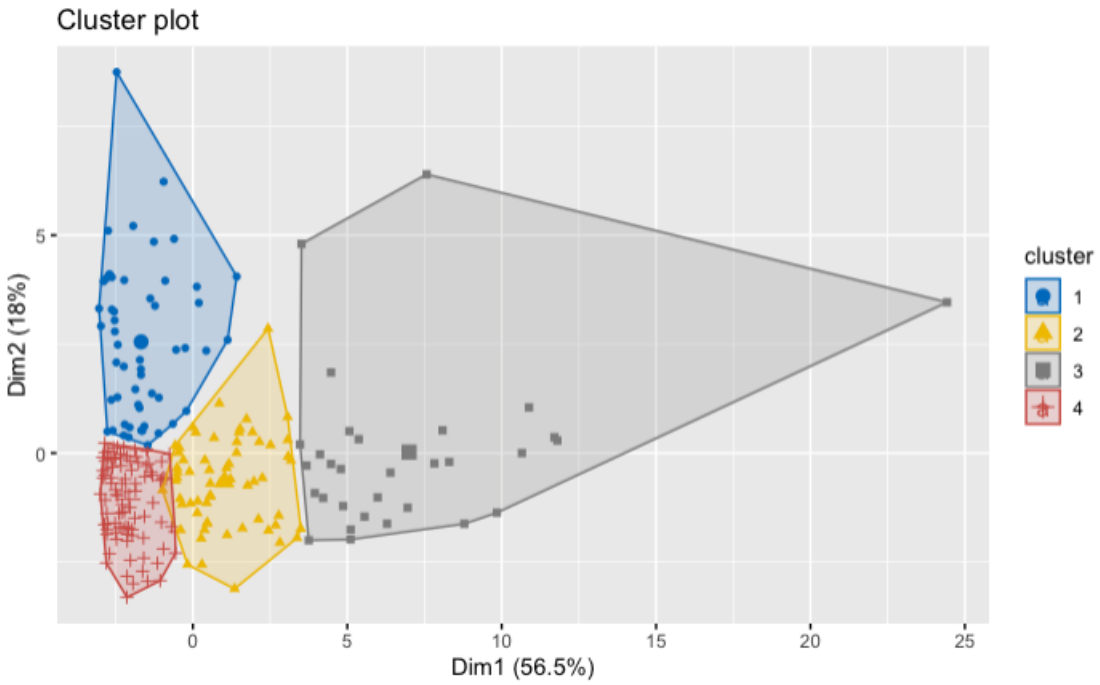
Description: df [5 × 5]

f0 <chr>	C1 <dbl>	C2 <dbl>	C3 <dbl>	C4 <dbl>
cEXT	0.09259259	0.41935484	0.45161290	0.51086957
cNEU	0.51851852	0.29032258	0.32258065	0.17391304
cAGR	0.20370370	0.22580645	0.00000000	0.00000000
cCON	0.12962963	0.01612903	0.06451613	0.21739130
cOPN	0.48148148	0.03225806	0.16129032	0.04347826

5 rows



Cluster 1 has a higher comparative level of neurotic and insightful users, which is 51% and 48%. Cluster 2 contains 41% sociable users, and very few conscientious users at only 1.6%. Cluster has 45% sociable users and no friendly users. Cluster4 has 51% sociable users and no friendly users, which is similar as cluster 3. With several clusters having similar proportions, it doesn't seem that appropriate distinctions can be made. It seems that cluster 3 and cluster 4 have some overlapping. However, according to the cluster plot diagram, the clusters which are overlapped together are actually cluster 2 and cluster 4. We could also observe this in the cluster plot diagram. It is also a contradiction that based on the cluster 3 and cluster 4, sociable people are uncooperative people. Therefore, when we compare with the K-means clustering results, we don't think that our fuzzy C-Means gave clearly defined results.



4 Conclusion and Discussion

To conclude, we started with a large dataset with almost ten thousand Facebook statuses from 250 users. We added several new aggregate level features and several deeper level features defining a user's emotions. Before adding the emotional features, our basic K-Means model did not provide very useful information. However, after accounting for new features like "anger," "trust", and so on, we our models outputted clearly defined clusters for different types of personalities. Our models performed the best at predicting either highly conscientious or highly neurotic personality types. This suggests that there are certain clear differences in the textual composition of the user's posts who were either neurotic or conscientious. We also may estimate that there is some likelihood of a difference in emotion that can be found between conscientious and neurotic Facebook statuses.

In further research, we would look at higher level natural language processing models, such as parts of speech and bag-of-words.

5 Reference

Chmielewski, M.S., Morgan, T.A. (2013). Five-Factor Model of Personality. In: Gellman, M.D., Turner, J.R. (eds) Encyclopedia of Behavioral Medicine. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-1005-9_1226

Sutin, A. R., Terracciano, A., Deiana, B., Naitza, S., Ferrucci, L., Uda, M., Schlessinger, D., Costa, P.T., Jr (2010). High neuroticism and low conscientiousness are associated with interleukin-6. Psychological medicine, 40(9), 1485–1493. https://doi.org/10.1017/S0033291709992029

LeMonda, B. C., Mahoney, J. R., Verghese, J., Holtzer, R. (2015). The Association between High Neuroticism-Low Extraversion and Dual-Task Performance during Walking While Talking in Non-demented Older Adults. Journal of the International Neuropsychological Society : JINS, 21(7), 519–530. https://doi.org/10.1017/S1355617715000570