Personality Classification with Social Media Group 29

Raahul Vignesh Manikandan
rmanika1@asu.edu
1219465754

Madhura LaalasaVallika Ganga
mganga@asu.edu
1220012365

Bhavani Mahalakshmi Gowri Sankar
bgowrisa@asu.edu
1222186719

Sanjeev Ganga Raju
sgangar2@asu.edu
1212096117

Sharath Srikanth
ssrika14@asu.edu
1219430706

Digvijaysingh Rajput
dsrajput@asu.edu
1219750766

***Abstract*—Recent years have seen an increase in the usage of online social networks and it has not only become a place to identify oneself, but also to curate, and create content. This emergence of social media has led to a massive increase in online data and data-driven computing techniques like machine learning, artificial intelligence, and data science. Using the available data and such modern techniques, traditional problems are starting to be solved in a data-centric manner. Similarly, in the field of behavioral learning and human psychology, traditionally, entities were only able to access an individual's personality through having them fill out psychological questionnaires. However, with the availability of data from the social media networks we could employ machine learning techniques to predict a user's personality directly. Information on users and what they communicated through their status or reactions to various posts is quite valuable and capturing the intricacies of this data could help us measure a person's personality in a variety of methods, including the Jungian personality test, Freudian analysis, Myers-Briggs analysis and the Big Five Personality test. In this project, such an attempt has been made to predict the personality of an user using the data extracted from Facebook (Currently know as Meta) and employing the Big Five Personality framework. Thus, the main objective of the paper focuses on employing Machine learning models to predict the personality traits and eventually, use this data to estimate the leadership abilities and emotional quotient of a user.**

***Index Terms*—Personality prediction, Big Five Model Personality, Facebook, Machine learning**

I. Introduction

Digital products and services are increasingly being used to mediate human activities (Lambiotte Kosinski, 2014). Individuals converse using social networking sites and messaging apps, make payments using online platforms and credit cards, and stream digital media utilizing online platforms and credit cards. In addition to that, the daily activities of the user are being captured by wearable devices like smartphones, fitness trackers, smart watches, and many more. Moreover, web browsing logs, transaction records, online articles, status, tweets, snap streaks, images, videos, media playlists, video call transcripts, emails and more more, have been an integral part of daily activities and have resulted in majority of the online data in the last decade. This torrent of user data clubbed with massive computing power and modern data-driven statistics and techniques, has revolutionized major departments in science including the social sciences and human psychology. Moreover, using computer based techniques like machine learning has helped the domain to capture even the minuscule of knowledge, thus resulting in more accurate as well as data-supported decisions. Therefore, the usage of such modern statistical tools and techniques from Data Science and Machine Learning techniques have become quite common. As a result, the traditional problems present in these domains have been redefined and restructured to use these techniques to solve the problems in better fashion.

Personality prediction is a traditional problem that is in existence for quite some time now. Traditionally, this problem has been solved by asking users to fill-out a psychometric questionnaire and manually analyzing these answers to figure out the personality. Such a task is quite demanding and time consuming as it involves more human-based jobs than employing computational facilities. This problem of predicting a user's personality in a faster yet more accurate way has been a question of research in the field of human psychology. As a result, employing Machine learning models and modern statistical tools to solve this research is a compelling necessity. Moreover, to make such an attempt, it is important to use the theories from the targeted domain in-order to facilitate an appropriate transfer of field knowledge. Thus, to predict the personality of a user, one should be aware of the frameworks in psychology that can aid in this process. One such theory that is being used in this paper is the Big Five personality Trait model. Also popularly known as the OCEAN model, this psychological theory is proven to estimate the overall personality of a person using five traits. They are Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness(A), and Neuroticism (N).

- **Openness**: Associated with openness to creativity, experience, also thinking of the person, and insight.
- **Conscientiousness**: Includes thoughtfulness, good impulse control, goal-directed behaviors, level of organization, and sense of duty.

- **Extraversion**: Characterized by an individual's social skills, also displays an active and assertive personality.
- **Agreeableness**: This showcases a sense of trust, altruism, kindness, and affection.
- **Neuroticism**: Linked with depression and insecurity, sadness, moodiness, emotional instability, and anger.

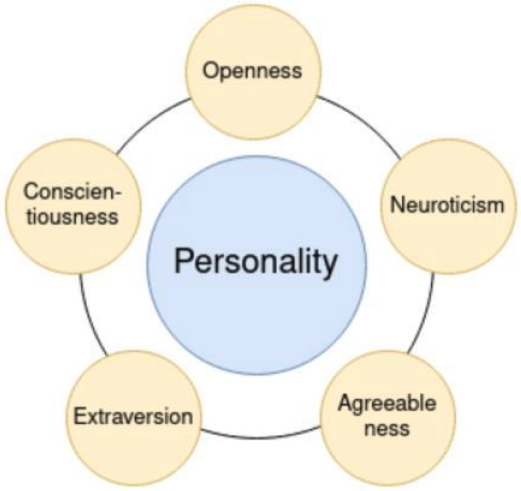


Fig. 1. Demonstrating the five traits of a personality[3]

Using this OCEAN model as the defacto standard for personality prediction, we are aiming to create a Machine learning model that can take a pre-processed user data and give out the leadership quality of an user. Here in this project, we have used the dataset extracted from one of the largest social media platforms, the facebook. The more in-depth summary on the used datasets is provided in the dataset section. While there are multiple personalities that a human exhibits, one of the crucial personality traits is leadership. Measuring this ability is not only non-trivial but also spans a wide range of applications from job hiring to political campaigning and many more.

An ideal leader is someone who is helpful, organized, confident, and social. These traits are directly related to the big five personality traits of the OCEAN model. A person who possesses favorable leadership qualities scores high in the openness and conscientiousness categories, average in the extraversion and agreeableness categories, and low in neuroticism [4]. In addition to the big five personality traits, the emotional quotient of an individual is an important attribute in terms of leadership. Leaders are frequently called upon to empathize, relieve stress, and solve conflict which requires emotional understanding and control. In order to "score" individuals on the OCEAN model and measure their emotion, one must collect data related to a user's personality. The rapid growth in social media usage and data availability makes a user's social media activity the ideal candidate for personality prediction. This project will analyze a user's Facebook likes to predict the OCEAN values of the user and their emotional quotient. These two calculations will then be used to score the user's leadership ability/strength.

II. Problem Statement

There are more than 4.55 billion active users on Social media as of October 2021 according to a report published in [2]. The amount of information originating from these users engaging and reacting to the content available on these social media platforms is fascinating. This novel information can be vital in providing valuable insights into users' personalities without formally taking a personality test. This approach, involving the extraction of digital content and drawing the personality model based on it is known as Personality Prediction.

We focused on predicting the user's personality traits and leadership abilities. We used the big five personality traits model for personality evaluation. It is renowned for its simplicity and verified proficiency. We also aimed to calculate the user's emotional quotient. We sought to get a better understanding of the user's awareness about handling his or her emotions. Also, it will provide better discernment into how well the user is able to overcome obstacles, communicate, relieve his stress and most importantly empathize with others. Lastly, we intended to discover the user's type of leadership based on the results.

III. Related Works

Personality prediction has become a topic of research for a while now, especially with abundant social media users. The prediction can be made based on different attributes, for example a tweet made by a user, images posted, and users likes. In [5] they have mentioned that personality prediction can be done from social media texts, as it's used to express themselves. This paper also talks about the importance of automatic personality prediction like efficiency and accuracy. They have used DISC benchmarks for assessment as it concentrated on behavioral preferences. In [6], twitter data was used and personality assessment models. They used Zemberek-NLP to get the words and LIWC to compute the frequency of words in text. It has the ability to detect individual differences and thinking styles. In [7] they are predicting personality based on Facebook user likes. Personality is classified to Big five traits, so they train their machine learning models with facebook's hierarchy of pages. Then use classifiers like neural networks and boosted trees to get individual personality scores. We used k-nearest neighbors and regression techniques to get the scores and finally arrive at an emotional state of a user.

IV. System Architecture Algorithms

The system architecture constitutes the success of the entire application. We began by preprocessing the data by eliminating the users and the likes that were of less than the threshold value. The reason for doing this is to eliminate the users and likes that would otherwise act as an outlier. The prediction model built for this dataset would not perform well. As a next

step, we built a sparse matrix from the users table to perform clustering[1]. We have used SVD to form the clusters. PCA was a good alternative to do the task, however, due to the time complexity involved in it, it was not feasible to do it. On the other hand, SVD is a matrix factorization technique that could be used to reduce the dimensions and form clusters. Once we have done this, we need to find an optimal value k for the SVD that contributes to the SVD dimension. Hence, we check for various values of k from 2 to 50 and train the model to check the prediction accuracy for the five personality traits. We are using a linear regression model to train the model and predict the values. The optimal k value was found to be 50 for the process. Hence, once we know the optimal k value, we build the predictor model to predict the personality traits for every user. We had the leadership personality traits weights precomputed based on the different websites and research articles we went through. To predict the leadership abilities of every user, we have used the cosine similarity to compare the given leadership weights with the predicted user personality traits. The resultant scores tell how much of a leader he/she is [-1,1]. As a final step, we calculated the emotions for the users in the following manner. We first found out the emotions associated with every like from the dataset. Next, we computed the aggregated emotions associated with each user to understand which emotion contributed to the maximum value among all the other emotions for a particular user. Once we have done this, the resultant table would give us the users and their associated emotions based on the posts liked by them. We have made use of a R package called Syuzhet to accomplish the task. In this way we were able to calculate the leadership abilities and the emotion quotient for every user.

We have used many algorithms in the project. To start we tried using PCA which stands for Principal Component Analysis [8] which is a dimensionality reduction technique. It is computationally expensive as it involves matrix multiplication. Hence, we moved to SVD which stands for singular value decomposition which is a matrix factorization technique which is efficient and computes the values in a faster way. It generalizes the eigen decomposition of a square normal matrix with an orthonormal eigenbasis to any (m x n) matrix. It is related to polar decomposition [9]. To train the model, we have used the linear regression model. [10] By reducing the sum of the squares of the vertical deviations from each data point to the line, this method calculates the best-fitting line for the observed data (if a point lies on the fitted line exactly, then its vertical deviation is 0). There are no cancellations between positive and negative numbers since the deviations are squared first and then summed. To calculate the leadership abilities, we have used the cosine similarities [11]. Finally the emotion quotient is evaluated by making use of the Syuzhet package in R.

V. Dataset

The data set used for this project was created by David Sitwell and Michal Kosinki and is based upon users and their Facebook likes [3]. The set contains three files: users.csv, likes.csv, and user-likes.csv. The users.csv file contains the userId, gender, age, political view, and OCEAN values for 110,728 users. Likes.csv has the likeId and like text of 1,580,284 posts. The user-likes contains the userid to likeid combination for 10,612,326 likes. By having all this data in this structure, it allows us to find all the likes of a user and the contents of each like's post.

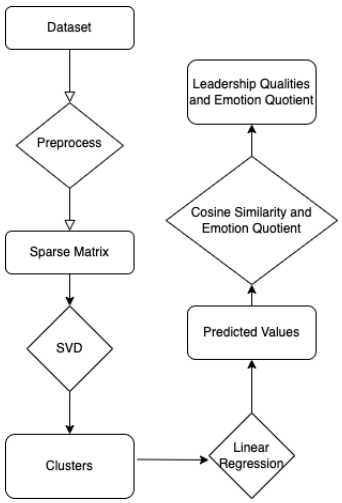


Fig. 2. System diagram.

Preprocessing consisted of loading the csv files into dataframes and dropping unnecessary data. Only the political value column was dropped in the users table because more than 75 percent of the values in the column were null or empty. Next, several joins were performed to get all the likes for each user and the text of each like. This was accomplished by joining the user table with the userLikes table based on userid. The resulting table contained the user information combined with the corresponding likeid. Next, this table was joined with the likes table based on the like id which resulted in every row containing a userId, user details, likeId, and the like text for every userId and likeId combination. With this information, we then moved on to build our prediction model and evaluate it in order to predict the OCEAN values.

VI. Evaluations

To train the model in the best possible way, we need to choose the best k value so that we form the right clusters. To check this, we ran the code by trying various k values and observed that k = 50 was the best value for which we were able to achieve the highest prediction accuracy.

The figure above shows the prediction accuracy for the linear regression model we achieved for the 5 personality traits. The accuracy achieved is pretty amazing considering the fact that we have only made use of the user's likes. This shows that there is a strong co relationship between user likes and personality traits. The figure below shows the leadership qualities (value) predicted for each user.

The values are between [-1,1] denoting how much of a leader he/she is. We were fascinated by how we could predict

```
[1] "k-val 50  Variable ope done."
[1] "k-val 50  Variable con done."
[1] "k-val 50  Variable ext done."
[1] "k-val 50  Variable agr done."
[1] "k-val 50  Variable neu done."
$ope
[1] 0.4495968

$con
[1] 0.2639683

$ext
[1] 0.3087149

$agr
[1] 0.212923

$neu
[1] 0.3040935
```

Fig. 3. Cross-validated predictions for the particular fold number=10 and k=50

```
$d74b66a9208973ba0e396b8ca5c80d3f
          [,1]
[1,] 0.4039556

$`1111b964aa1d7bcac0cb562d1bfdce63`
          [,1]
[1,] 0.1743463

$`1e08f5392dff3d5024ee23886443bef2`
          [,1]
[1,] -0.2579741
```

leadership abilities with the given data. In the next step, we predicted the emotions associated with a user based on the posts liked by them. We took the skimmed likes table and for each like, we get a list of words pertaining to the like. Then, we pass each word obtained to an R package that gives us the NRC sentiment of the word. The result obtained consists of a data frame with the following emotions: ”anger”, ”anticipation”, ”disgust”, ”fear”, ”joy”, ”sadness”, ”surprise”, ”trust”, ”negative”, ”positive.”

like_sentiments

likeid	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	negative	positive
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
220f40e9d4e1bd4ec84d90fe7c13319c	0	0	1	0	0	0	0	0	1	0
34af92763e9adbf5f4a9160f200c11c2	0	0	0	0	0	0	0	0	0	1
6a7da93fd21d77eb40182cb1d7f698a2	0	0	0	1	0	1	0	0	1	0
a68fe60f5f469b442dee190151f8d63f8	0	0	0	0	0	0	0	0	0	0
5173dc869ee91c8f927d3564ca4c012e	0	0	0	0	0	0	0	0	0	0
4294a8290c90ebf0f65ec7c42b4e9b5f	0	0	0	0	0	0	0	0	0	0
924c72f8cc894b3742d56c925542fc9d	0	0	0	0	0	0	0	0	0	0
c5e1ae6ad374b8953de77779e1936257	0	0	0	0	0	0	0	0	0	0
e2bf997a5154e633c6b765e5c51b756c	2	0	2	1	0	1	0	0	2	0
bba6341ca8e144623873be8fb7ea274a	0	0	0	0	0	0	0	0	0	0
159d47e1718dffaa079a56595780172O	0	0	0	0	0	0	0	0	0	0
99ea24d3b1fb8787ae5d9b622602e4a1	0	0	0	0	0	0	0	0	0	0

The figure above shows the output of the likes and its corresponding sentiments Next, the 10 values for each word are combined for a like and a data frame is created with the corresponding like ID and the emotions. The same process is repeated for all the likes in the likes table.

The picture above represents the user and his/her corresponding NRC sentiment values The next step was to take the skimmed list of users that we acquired earlier and use it to skim the user-likes table of the dataset. This way, we only get the user-like entries of the users and likes that we are currently processing. For each user, a list of likes is identified and for each like in the list, its corresponding 10 sentiment is added. The same process is repeated for all the likes of the user and a new data frame is created pertaining to the user ID

user_sentiments

userid	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	negative	positive
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
aae7df9b525a51bc2b073c4c96a214e6	0	0	1	0	0	0	0	0	1	0
e16ed09edcaf821b5a94e62bdd6094eb	0	0	0	0	0	0	0	0	0	1
dd25e93bf2ad2d7faaaba02e807728b4	0	0	0	1	0	1	0	0	1	0
2eebab587655fd7c013ad75e32c0c8a9	0	0	0	1	0	1	0	0	1	0
ff8629b5470c6285bc1a6540f0cc0ec6	0	0	0	0	0	0	0	0	0	0
bbe5bc49f00dea56af003f6c32d430e3	0	0	0	0	0	0	0	0	0	0
feaafa395a3a16e1c00c6afacc24b81d	0	0	0	0	0	0	0	0	0	0
d26d7624ef925a9a5ced1a96be423020	0	0	0	0	0	0	0	0	0	0
a4989234dd833a40639aed911e35ec45	0	0	0	0	0	0	0	1	0	1
3f4cfca8dd5a285809d093d34db142c9	0	0	0	0	0	0	0	0	0	0
ef62ea9b5993c56f65226b29ddfc0794	0	0	0	0	0	0	0	0	0	0

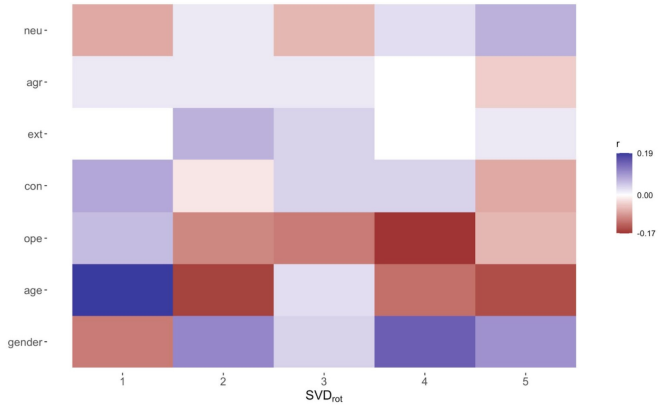
and the emotion values depicting the sentiments of the user. This part of the code requires heavy processing of the data and hence it was executed on Google Colab.

	userid	sentiments
	<chr>	<chr>
1	aae7df9b525a51bc2b073c4c96a214e6	disgust negative
2	e16ed09edcaf821b5a94e62bdd6094eb	positive
3	dd25e93bf2ad2d7faaaba02e807728b4	fear sadness negative
4	2eebab587655fd7c013ad75e32c0c8a9	fear sadness negative
5	ff8629b5470c6285bc1a6540f0cc0ec6	anger anticipation disgust fear joy sadness surprise trust negative positive
6	bbe5bc49f00dea56af003f6c32d430e3	anger anticipation disgust fear joy sadness surprise trust negative positive
7	feaafa395a3a16e1c00c6afacc24b81d	anger anticipation disgust fear joy sadness surprise trust negative positive
8	d26d7624ef925a9a5ced1a96be423020	anger anticipation disgust fear joy sadness surprise trust negative positive
9	a4989234dd833a40639aed911e35ec45	trust positive
10	3f4cfca8dd5a285809d093d34db142c9	anger anticipation disgust fear joy sadness surprise trust negative positive

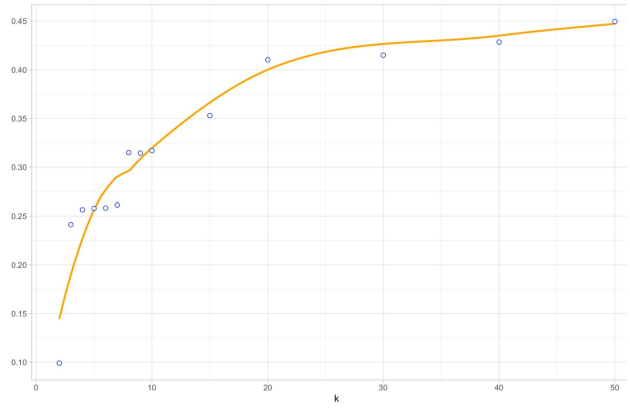
In this way we were able to calculate the personality traits, leadership abilities using those predicted personality traits and finally the emotions associated with the users as seen above.

VII. Visualizations

The heat map shows the results of the correlation between the attributes of the users table. We can clearly observe that openness has the highest correlation among all other attributes. Hence we tried to build a model and check the prediction accuracy against openness which would give us a fair amount of knowledge about the model.



The figure below shows the prediction accuracy for various values of k. This plot helped us to choose the best k value for the SVD technique.



VIII. Division of Work and Team Members' Contributions

There were three main phases during the project: Research/Planning, Building/Evaluating, and Reporting/Presenting the project. During research and planning Sanjeev and Raahul were in charge of looking for potential datasets and libraries that would help in building a personality prediction model. Sharath and Bhavani were in charge of researching useful algorithms and model architecture while Digvijay and Madhura were in charge finding research papers and similar projects on personality classification. The division of tasks for building and evaluating the model can be found in the table below.

TABLE I
Division of Work

Task	***Members Assigned***
Data Collection and Pre-processing	Sanjeev/Raahul
Research/Find clusters	Sharath/Bhavani
Perform clustering/dimensionality reduction	Digvijaysingh/Madhura
Build prediction model	Sharath/Bhavani
Evaluate model and results	Sanjeev/Raahul

When writing reports and presenting, all members worked equally and collectively to write the document and create the presentation.

IX. Conclusions

We have shown that personality and emotions can be predicted with Facebook data. The dataset we had was satisfactory in terms of results we have seen. Like discussed several 'K' values were tested for SVD clustering and we got promising results when K was set to 50. Out of all the Big five traits, Openness has highly correlated with attributes in data(as seen in heat map above). We could define emotional states of a user like happy, sad, anger etc. Therefore, user personality and emotions can be predicted from their social media footprint. Future work includes training the model with other classifiers by implementing LDA or k-means clustering. Also, training with different attributes like images and texts collected from the user's profile.

References

[1] https://www.michalkosinski.com/data-mining-tutorial
[2] https://wearesocial.com/jp/blog/2021/10/social-media-users-pass-the-4-5-billion-mark/
[3] https://www.enjoyalgorithms.com/blog/personality-prediction-using-ml
[4] https://sites.psu.edu/leadership/2019/05/16/leadership-with-the-big-five-personality-traits/
[5] P. S. Dandannavar, S. R. Mangalwede and P. M. Kulkarni, "Social Media Text - A Source for Personality Prediction", 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)
[6] İzel Ergu, Zerrin Işık and İsmail Yankayış, "Predicting Personality with Twitter Data and Machine Learning Models", 2019 Innovations in Intelligent Systems and Applications Conference (ASYU)
[7] Raad Bin Tareaf, Seyed Ali Alhosseini, Philipp Berger, Patrick Hennig and Christoph Meinel, "Towards Automatic Personality Prediction Using Facebook Likes Metadata", 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)
[8] https://en.wikipedia.org/wiki/Principal_component_analysis
[9] https://en.wikipedia.org/wiki/Singular_value_decomposition
[10] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm
[11] https://www.sciencedirect.com/topics/computer-science/cosine-similarity