Personality Classification with Social Media

Kunj Patel
ASU ID: 1213184152
khpatel8@asu.edu

Sreshta Chowdary Kampally
ASU ID: 1224031900
skampall@asu.edu

Sai Ajitesh Krovvidi
ASU ID: 1219320388
skrovvid@asu.edu

Phani Rohitha Kaza
ASU ID: 1219915970
prkaza@asu.edu

Prakruti Singh Thakur
ASU ID: 1222301340
psthakur@asu.edu

Ankit Sharma
ASU ID: 1219472813
ashar263@asu.edu

***Abstract*—Social media is the largest platform, with over 4.62 billion users from diverse demographics and backgrounds. These users actively express themselves and share their opinions and feelings in the form of text or images from time to time. These opinions and feelings, along with the person's attitude, likes and dislikes, and various other characteristics, constitute a person's personality. Personality is an essential component of an individual's perceptual and intuitive behavior, making it important. Currently, there are multiple ways to classify a person's personality. The Big 5 factor is one of them. It is usually referred to as the OCEAN traits [7] and can be defined as follows:**

S.No.	Trait	Description
1	Openness	Represents the ability to be open to new experiences and take part in imaginative, out-of-the-box activities
2	Conscientiousness	The ability to show high levels of thoughtfulness, good impulse control, and exhibit goal-directed behaviors
3	Extraversion	Exhibit high energy, socializing behavior, assertiveness, and high amounts of emotional expressions.
4	Agreeableness	Characterized by trust, altruism, kindness, affection, and other prosocial behaviors.
5	Neuroticism	Characterized by sadness, moodiness, and emotional instability

Index Terms*— *Facebook, Personality Classification, OCEAN traits, Clustering, KMeans, DBSCAN, Fuzzy C-Means, Latent Dirichlet Allocation, Singular Value Decomposition

I. Introduction

In this paper, we discuss how user personalities can be classified through their social media data - their gender, age, political views, personality traits and posts liked. There are multiple ways in which a person's personality is classified, but we will focus on the "Big Five" model in our paper. It states that personalities can be widely classified as Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. In the last few years, there has been a lot of research in the field of personality prediction through social media. It's because it can be used to enhance users' experience to personalize social media and e-commerce content for targeted sales, workforce selection, planning, etc.

In order to classify and predict these personalities, we collected Facebook user data. The data collected is huge in number and has over 110,726 unique Facebook users and the different posts they have engaged with by 'liking' them. The various observations we made looking at the data were:

- 44 percent users were Male and 66 percent were Female
- "Neuroticism" was the most expressed trait and "Openness" was the least among people (especially, in females)
- a higher percentage of men exhibit "Conscientiousness", "Extraversion" and "Openness" traits than females

Next, we pre-processed our data in order to generate a user-footprint matrix which contained the users and the data associated with their likes. We tend to remove the data where the user and likes occur very sparsely.

Our aim is to build an Unsupervised learning model for predicting a user's personality based on personality traits, age, gender, political views, and pages liked. The first step taken was, to perform Dimensionality Reduction on this large data in order to prevent overfitting and to get rid of noise and false correlations. Particularly, we used to Singular Value Decomposition (SVD), Principal Component Analysis (PCA), and Latent Dirichlet Allocation (LDA) to identify which method is best for our data.

Next, Clustering models K-Means and DBSCAN were used to group these Facebook users depending on the posts they liked. It was performed on the resulting user-footprint matrix from data pre-processing. We also performed Linear and Logistic Regression to make predictions. In the further sections, we will be discussing in details the steps mentioned above.

II. Problem Statement

With the increasing usage of social media in recent times, it has been possible to get hold of a large amount of data. People interact on social media in many ways. Some of the most popular ways are through liking someone's posts or commenting on them. In the Q4 of 2021, the number of daily active users on Facebook reached 1.93 billion. With such a high volume of data availability, it is becoming more and more critical to classify an individual's personality traits to better serve their demands. At the same time, it could also translate to more revenue for the company by running targeted ads and marketing campaigns. This is also important because some

personality types are more likely to engage in certain types of activities. Knowing this, for instance, psychologists can better serve a patient's condition. It is clear that personality impacts behavior and lifestyle choices.

Given Facebook users' activity data in their likes, we want to classify their personalities based on the OCEAN principle. This can be done by applying state-of-the-art clustering algorithms that will get us high accuracy. There is already much research in this area, and if time permits, there's potential to extend this project to work for any kind of dataset without having access to specific types of social media activity attributes.

III. Related Works

With a large amount of data being collected by social media websites like Facebook and Twitter, there have been studies linking user behavior to their personality. The "Big Five" personality traits were proposed by Tupes and Christal [9]. The five factors identified based on lexical analysis for the US English-speaking population are [10] [11] -

- Openness - inventive, curious, consistent, cautious.
- Conscientious - efficient, organized, extravagant, careless.
- Extraversion - outgoing, energetic, solitary, reserved.
- Agreeableness - friendly, compassionate, critical, rational.
- Neuroticism - sensitive, nervous, resilient, confident.

These labels are typically referred to by using their acronyms "OCEAN" and "CANOE". This personality model is regarded as one of the most widely accepted and well-researched model. There have been previous studies to classify user personality into one of the ocean traits by collecting their interaction data with a social media website. For the scope of this project, we limit our discussion to three different research works [1] [5] [6], as these works are closely related to the goal of our project.

[5] Golbeck, Jennifer, et al presented a method to predict a user's personality by using the publicly available information available on their Twitter profile. The authors of the paper selected 50 subjects from Twitter, Facebook, and mailing lists. Users were then given a 45-question test based on the Big Five Personality test and then the most recent 2,000 tweets of each user were collected. Then the Pearson correlation analysis was done between every feature obtained from tweet analysis and the user's personality score. Correlation results were intuitive in some cases but did not provide explanations for some. Scope for further work was left for identifying these unexplained correlations over a bigger dataset. The prediction analysis was done using Gaussian Process and ZeroR regression analysis algorithms. This approach yielded expected results for openness and agreeableness but less accurate results for conscientiousness, extraversion, and neuroticism. The authors used the number of followers and network density of the subject as the features for generating personality scores. Using personality scores of followers, connection strength to a follower, and other related factors to improve the prediction results for personality classification were left for the scope of further research.

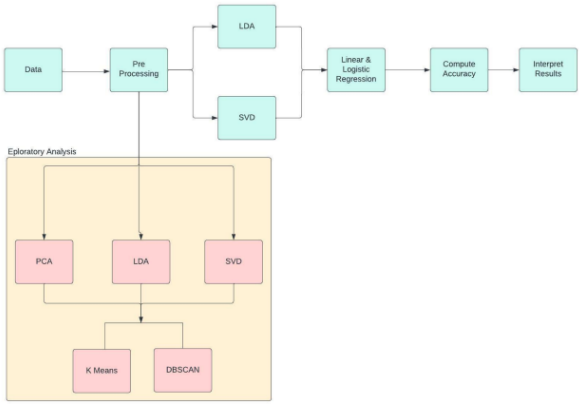

Fig. 1: System Architecture

[6] Quercia, Daniele, et al studied the relationship between OCEAN personality traits and different types of Twitter users, which included users with a large number of followers, influencers, and active users. The authors used myPersonality on Facebook and filtered 335 users of interest who had provided links to their Twitter profiles. These users were classified into listeners, popular, and highly-read based on their following, followers, and listed counts available publicly. Regression analysis was used for predicting personality and the authors were able to achieve a maximum root-mean-square-error (RMSE) of 0.88 between the predicted and observed values. The research study inferred that popular users tend to be imaginative, influential users organized, and all user types were low in Neuroticism and most were extroverts.

[1] Kosinski, Wang, et al paper provides an algorithmic framework and required tool which can be used to analyze user data gathered from social media and draw insights and build predictive models for real-life outcomes. Authors suggested that social media datasets contain traces of human behaviors and methods can be employed to extract patterns and build predictive models. Big data sets which provide users' digital footprints are represented as a user-footprint matrix where the rows represent the users and columns represented the digital activity of the user. Associations below a certain threshold are removed and the dimensionality reduction techniques like Singular Value Decomposition (SVD) and Latent Dirichlet Allocation (LDA) are used to reduce the dimensions and improve the accuracy of the predictive models. SVD and LDA are popular dimensionality reduction techniques often used in context of social sciences, natural language processing, and study patterns in languages. The strength of associations between the user-footprint matrix and clusters are used to interpret dimensions and clusters obtained after the dimensionality reduction. Authors suggest use of techniques such as linear and logistic regression which offers accuracy similar to advanced techniques. For future works, the authors suggest researchers to be wary of the methods used to obtain data because of the vague boundary between public and private information.

IV. System Architecture & Algorithms

In this section, we focus on discussing our project's system architecture in detail. A high-level overview of the project's architecture is shown in figure 1. We will explain the associated techniques, algorithms, and processes used in each step of the architecture.

A. Data

The dataset obtained for this project was downloaded from the companion website (http://datamining tutorial.com) of [1]. It contains three files - users.csv containing features for 110,728 users, likes.csv stores information about like IDs and its name, and users-likes.csv which stores associations between users and like.

B. Preprocessing

In the preprocessing step, a user-footprint matrix was generated to store data related to the user and the frequency of words associated with the like data which help to generate n-grams for each user. The user-footprint matrix was converted to a sparse matrix which helped reduce the storage requirement for the matrix. We also removed information from the sparse matrix which did not meet our threshold of likes per user and users per like metrics.

C. Dimensionality Reduction

The next step in our project was to extract patterns from the user footprint matrix. To extract patterns we first needed to reduce the dimensionality of the data because it is benefits like reducing the number of features than the number of users, lowering the risk of overfitting, removing the redundancy in data, making data more interpretable, and reducing the computation time and memory requirement in further steps. We used Singular Value Decomposition (SVD), Principal Component Analysis (PCA), and Latent Dirichlet Allocation (LDA) techniques as the main consideration for dimensionality reduction.

- SVD is popular in the area of social sciences and natural language processing. The optimum value of the number of SVD dimensions is selected from the 'knee' of the singular values vs k plot.
- PCA is a popular choice amongst researchers but it is computationally expensive as it requires multiplying a matrix transpose with itself.
- LDA has the advantage of ease of interpretation but it can only be applied to a non-negative dataset, which is true for our case.

D. Clustering

This step is important to interpret dimensions extracted from the user footprint matrix. This step provides insight into the data processing output generated and helps us guide in choosing techniques and algorithms for the next steps. We generated heatmaps for LDA and SVD clusters. The heatmaps generated with users classified into k=5 clusters ans their psychodemographic traits provided insights into information like:

- Young and female users liked humorous posts and interact more with Facebook advertisements.
- Mature and liberal users tend have likes associated with shows like The Colbert Show ans personalities like Barack Obama.
- Female and conservative users have likes associated with pop music.

In our exploratory analysis, we also performed clustering using DBSCAN and K-Means after reduing dimensionality with PCA, SVD and LDA algorithms. In this process we observed that reducing dimensionality used PCA and then performing clustering using K-Means provided with most well-defined clusters, as shown in figure 15.

E. Build prediction models

We used linear and logistic regressions approaches to build prediction models after performing dimensionality reduction and clustering techniques as discussed in the previous sections. Using linear and logistic regression provided us the benefit of being fast and simple. To avoid overfitting our model we divided our dataset into training and testing sets.

V. Dataset and Data Preprocessing

The dataset we are using has been obtained from the website "https://www.michalkosinski.com/datamining-tutorial." This data contains three files which contain the profile information and personality traits, as well as the data that includes the associations between the users and likes, from Facebook.

- **user.csv:** This file contains the data of 110,729 Facebook users. It includes the userids and their respective gender, age, political inclination, and 5 personality traits.
- **likes.csv:** This file contains 1,580,284 likeids and the name of the post that is liked.
- **users-likes.csv:** This file contains 10,612,326 lines of data. This data comprises two columns which are userid and the likeids of the posts they liked.

VI. Evaluations

We compared the accuracies of both the model generated using SVD and LDA as shown in figure 6. We observed that for Openness, Conscientiousness, Extroversion, Neuroticism, Agreeableness, Gender, and Political factor - SVD had better accuracy. But for Age, LDA gave better results as shown in the table in figure 2.

In our exploratory data analysis, for the Facebook dataset with 110,729 users of which 44% are Males and 56% Females, we also observed following interesting metrics as represented in figure 3 -

- **Neuroticism** was the most expressed trait among people (especially, in females).
- **Openness** was the least expressed trait (especially, in females).
- Both genders show the same amount of **Agreeableness**.

- Higher percentage of men exhibit **Conscientiousness**, **Extraversion**, and **Openness** traits than females.

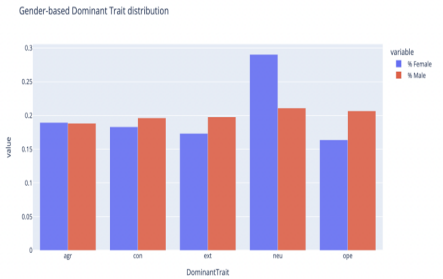

Fig. 2: Gender based Dominant trait Distributions

VII. Results & Data Visualizations

In this section we provide a visualization of correlations between feature clusters and different user traits, accuracies recorded for SVD and LDA models, and the clusters generated using DBSCAN and K-Means algorithms.

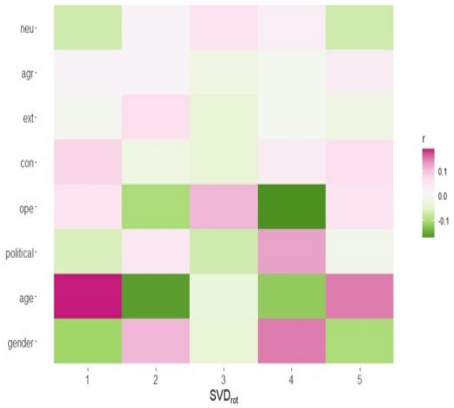

Fig. 3: Heat Map of SVD

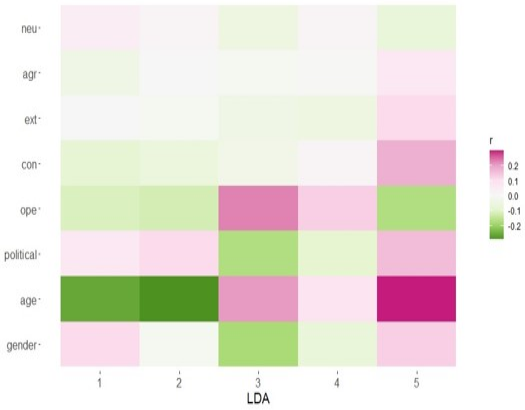

Fig. 4: Heat Map of LDA

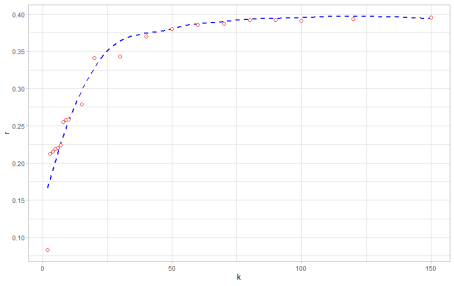

Fig. 5: SVD vs Dimensions Plot

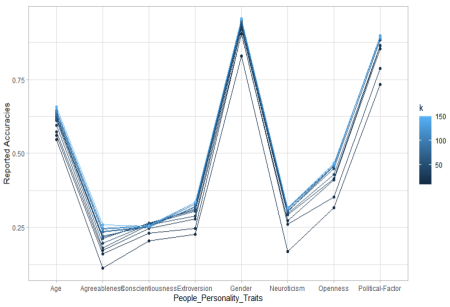

Fig. 6: Prediction Plots-Accuracies vs Traits(SVD)

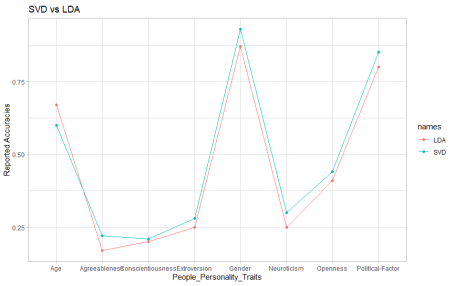

Fig. 7: SVD vs LDA Plot

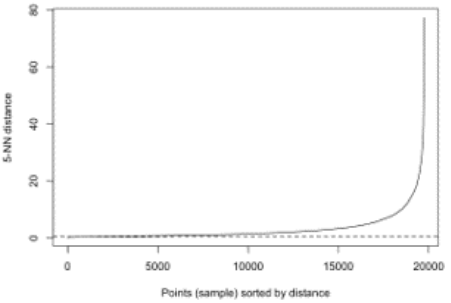

Fig. 8: DBSCAN PCA: SNN-Distance vs Points(sample) sorted by distance

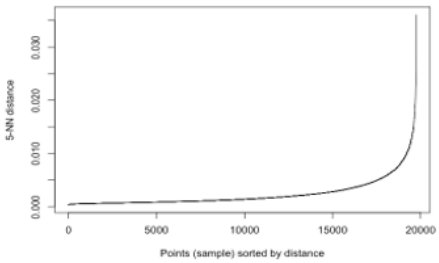

Fig. 9: DBSCAN SVD: SNN-Distance vs Points(sample) sorted by distance

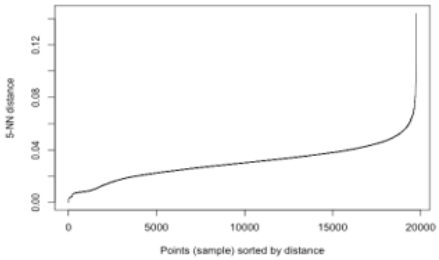

Fig. 10: DBSCAN LDA: SNN-Distance vs Points(sample) sorted by distance

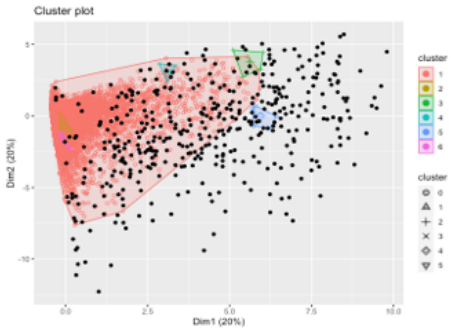

Fig. 11: DBSCAN: Cluster Plot for PCA

Fig. 12: DBSCAN: Cluster Plot for SVD

Fig. 13: DBSCAN: Cluster Plot for LDA

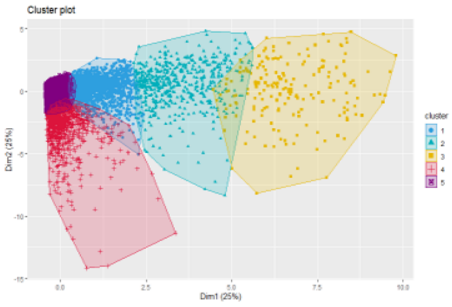

Fig. 14: K-Means: Cluster Plot for PCA

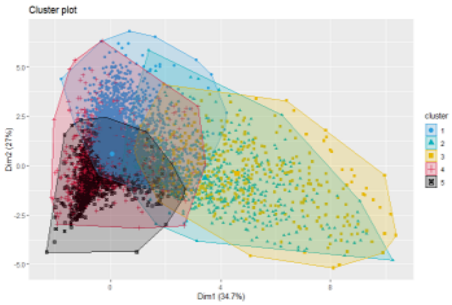

Fig. 15: K-Means: Cluster Plot for SVD

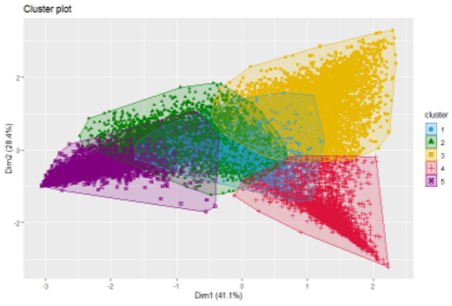

Fig. 16: K-Means: Cluster Plot for LDA

VIII. Division of Work and Team Members' Contributions

S.No.	Task	Assignee
1	Data Collection & Pre-processing	Sai Ajitesh Krovvidi, Kunj Patel, Prakruti Singh Thakur
2	Data Analysis & Visualization	Phani Rohitha Kaza, Kunj Patel, Prakruti Singh Thakur
3	Research on Dimentionality Reduction and Clustering Algorithms	Sai Ajitesh Krovvidi, Sreshta Chowdary Kampally, Ankit Sharma
4	Perform the Dimensionality Reduction and Clustering Algorithms	Sai Ajitesh Krovvidi, Sreshta Chowdary Kampally, Kunj Patel
5	Building the model for predictive Analysis	Phani Rohitha Kaza, Sreshta Chowdary Kampally, Ankit Sharma
6	Compare the results of the various Classification models and Choose the appropriate model	Phani Rohitha Kaza, Sreshta Chowdary Kampally, Prakruti Singh Thakur

IX. Conclusion

The amount of data that exists on social media is burgeoning due to large number of users using the platform for a wide range of purposes. Analyzing the traits of these users can be useful for multitude of reasons such as predicting political inclinations, running targeted marketing campaigns etc. During our project we got the opportunity to perform some exploratory analysis on the data set and we were able to figure out few interesting facts such as the users file contained a female to male ratio of 1.27:1. Also, male and female show the same amount of the Agreeableness trait. To begin with we performed cleaning and preprocessing on our data. Secondly, we reduced the dimensionality of the data by applying two types of dimensionality reduction techniques, viz. SVD, LDA to reduce false correlations. To continue, we made use of libraries in R such as varimax rotate inorder to obtain stronger correlations between the Eigen-traits and the dimensions. Later, Heat Maps have been produced for understanding the correlation values of the traits with the five dimensions.

The accuracy of multichotomous traits was measured using the pearson correlation coefficient and a graph corresponding to the value of 'r' versus the number of dimensions in SVD was plotted for a few traits to enhance the understanding of the readers. The number of dimensions that produced maximum accuracy for each trait was duly noted. Similarly, area under curve method has been utilized to compute the accuracy corresponding to dichotomous attributes. Finally, prediction plots for accuracies versus traits was plotted to analyze the results for prediction. We have carefully compared the results obtained by generating prediction models after using both LDA and SVD for reducing the dimensions and came to a conclusion that models generated after using SVD had made better predictions for all the traits except Age.

Beyond the scope of our project we have also used Principal Component Analysis along side LDA and SVD as a forestep to K-Means AND DBSCAN. From the clusters obtained we were able to come to a conclusion that we well defined clusters are observed when PCA was used as a dimensionality reduction technique before clustering. This is a pivotal step in comprehending more complex information such as political inclinations and we are one step away from obtaining all such information. Our group sees a lot of potential in this project and believe that further research can be made using our results as a reference to comprehend more complex information and produce higher levels of inference.

References

[1] Michal Kosinski, Yilun Wang, Himabindu Lakkaraju, and Jure Leskovec, "Mining Big Data to Extract Patterns and Predict Real-Life Outcomes", 2016.

[2] Michal Kosinski, https://www.michalkosinski.com/data-mining-tutorial, R code

[3] Dimensionality reduction techniques, https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b11-dimensionality-reduction-techniques-you-should-know-in-2021

[4] Clustering Algorithms, https://towardsdatascience.com/17-clustering-algorithms-used-in-data-science-mining-49dbfa5bf69a17 Clustering Algorithms Used In Data Science and Mining

[5] Golbeck, Jennifer, et al. "Predicting personality from twitter." Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE, 2011.

[6] Quercia, Daniele, et al. "Our Twitter profiles, our selves: Predicting personality with Twitter."Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third InernationalConference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE, 2011.

[7] OCEAN traits https://www.verywellmind.com/the-big-five-personality-dimensions-2795422the-big-five-personality-dimensions

[8] Statista https://www.statista.com/statistics/346167/facebook-global-dauFacebook DAU statistic

[9] E. Tupes and R. Christal. Recurrent personality factors based on trait ratings. Journal of Personality, 60(2):225–251, 1992.

[10] Sonia, Lilach, et al. "The Big Five Personality Factors and Personal Values". Sage journals, 2002.

[11] Big Five Traits, https://en.wikipedia.org/wiki/Big_Five_personality_traits