Project Proposal: Personality Classification with Social Media
Group 6

Kampally, Sreshta Chowdary
skampall@asu.edu
1224031900

Patel, Kunj
khpatel8@asu.edu
1213184152

Kaza, Phani Rohitha
prkaza@asu.edu
1219915970

Thakur, Prakruti Singh
psthakur@asu.edu
1222301340

Krovvidi, Sai Ajitesh
skrovvid@asu.edu
1219320388

Sharma, Ankit
ashar263@asu.edu
1219472813

March 02, 2022

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. 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. 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

Keywords: Facebook, Personality Classification, OCEAN traits, Clustering, KMeans, DBSCAN, Fuzzy C-Means, Latent Dirichlet Allocation, Singular Value Decomposition

1 Problem Definition

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.

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.

2 Data Sets

The dataset we are using has been obtained from the website "https://www.michalkosinski.com/data-mining-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

3 State-of-Art Methods & Algorithms

Accuracy for the user personality classification is mainly dependent on many factors at various levels ranging from pre-processing of the data to clustering techniques used. Firstly, it is essential to try to understand the data parameters that are quintessential for solving the problem. Reducing the dimensionality of our dataset using methods such as Principal Component Analysis (PCA), Latent Dirichlet Allocation (LDA), Singular Value Decomposition (SVD) helps in easier visualization of our data. Alongside avoiding the curse of dimensionality problem, it aids us in reducing false correlations, prevents overfitting, and, most importantly, removes irrelevant features for the scope of our project. The next step is to cluster the users; it is pivotal to use the proper clustering technique that suits our needs. Clustering techniques such as DBSCAN, K-Means that are most widely applied for similar problem statements will be researched to find out the most-apt algorithm that serves the purpose of our project. Clustering techniques can be compared against each other for performance by using various metrics such as the Silhouette score, Davies-Bouldin Index, and Calinski-Harabaz Index. After performing clustering based on the vital information, users will be assigned to groups and can later be utilized to classify a new user with the help of the models. Prediction models can be developed using Regression Techniques so on and so forth.

4 Research Plan

The techniques mentioned above for dimensionality reduction will be analyzed for applicability, and the method that produces the best results will be adopted. Similarly, for clustering, once we visualize our preprocessed data, we can predict the clustering techniques that deliver excellent results. Therefore, after initial observation, we plan to use the methods in the previous section for clustering. The clustering techniques will be compared for accuracy, and the more accurate method will be used moving forward. Analyzing the accuracy using multiple metrics will assist in arriving at sound conclusions. There are also numerous methods to build prediction models like linear and logistic regression, support vector machines (SVMs), neural networks, etc. Initially, we plan to use linear and logistic regression methods to build models and later work on sophisticated techniques such as SVMs and decision trees, considering time limitations. Finally, prediction models will be analyzed for accuracy to understand if a user's personality can be identified from the patterns obtained from our dataset.

5 Evaluation Plan

We plan on implementing various dimensionality reduction techniques with a combination of distinct clustering algorithms. Next, we'll be performing predictive analysis to predict the personality type from the OCEAN parameters. Upon comparing and analyzing the results for the mentioned models, we will use the model with the highest score for the chosen performance metrics. If time permits, we also plan to acquire the dataset for the other social media platforms like Twitter and work on it.

6 Project Timeline

S.No.	Task	Descriptions	Begin Date	End Date
1	Data Collection & Preprocessing	Collecting the data and preprocessing it from raw data to understandable format to the model	02-16-2022	02-25-2022
2	Data Analysis & Visualization	Analyse the data and have visualizations	02-26-2022	03-09-2022
3	Research on Dimensionality reduction & Clustering Algorithms	Conducting research to explore the best algorithms that suits our dataset in terms of reducing the dimensions of the data and clustering techniques	03-10-2022	03-17-2022
4	Perform the Dimensionality Reduction & Clustering Algorithms	Performing the Dimensionality Reduction & then Clustering Algorithms on the data	03-18-2022	03-25-2022
5	Building the model for predictive Analysis	Build the model for predictive analysis on the personality attributes	03-26-2022	04-07-2022
6	Compare the results of the various Classification models and Choose the appropriate model	Compare the results of various models and choose the appropriate one based on the performance metrics	04-08-2022	04-11-2022

7 Division of Work

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 & Clustering Algorithms	Sai Ajitesh Krovvidi, Sreshta Chowdary Kampally, Ankit Sharma
4	Perform the Dimensionality Reduction & 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

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, 11-dimensionality-reduction-techniques-you-should-know-in-2021

[4] Clustering Algorithms, 17 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 the-big-five-personality-dimensions

[8] Statista Facebook DAU statistic