Project Report : Personality Classification using Social Media
Group - 13

Amit Kumar Sinha
asinha58@asu.edu
1219500399

Padma Priya Patta
ppatta1@asu.edu
1219585965

Parimala Konduru
pkondur1@asu.edu
1219639538

Rama Lingeswara Naga Sai Kallakuri
rkallak1@asu.edu
1220421254

Rishitha Guptha Yedire
ryedire@asu.edu
1220826282

Sai Venkat Raavi
sraavi@asu.edu
1219639616

Vineeth Sai Surya Chavatapalli
vchavata@asu.edu
1218076795

Abstract

In this project, we explore Facebook as a data source for personality classification in Social Media. We present techniques below to determine the user's personality by utilizing the publicly available information about Facebook user profiles. Personality is the one which differentiates one person from another. The advancements in social media are coming out rapidly, increasing the human activities online. These activities indirectly reflect the behaviour or personality of a person. Social network investigators feel that studying the personalities of people has numerous advantages.

It is also possible to predict human personality based on their social media activities. Here we show that easily accessible digital records of behavior, Facebook likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender.

1 Introduction

There are multiple ways in which personality of a person is evaluated like, Jungian personality test, Freudian analysis and the Myers Briggs analysis. Out of all, The Big Five Model which is also known as the Five-Factor model is widely being used by psychologists. The model mainly depends on the five components of personality known by OCEAN.

- Openness: Indicates open-minded and authority challenging nature of a person
- Conscientiousness: Indicates self-disciplined nature of a person
- Extraversion: Indicates out-going and social nature of a person
- Agreeableness: Indicates warm,friendly and tactful nature of a person
- Neuroticism: Indicates a person’s ability to remain stable and balanced

2 Problem Statement

The large amount of digital records available to us on social networking platforms makes it possible to track and analyze meaningful traits and personalities. Now-a-days, every individual often interacts with social media through likes and comments which relate to their personality. With personality classification, we can foresee a person's likes and their interests. This classification can be advantageous in many ways like advertising and marketing purposes. In addition, this categorizations can be used by psychologists to perform critical analysis related to depression and suicide tendencies.

Among the questions we wish to address is actually whether we have to classify the character prediction process as a multi label one or perhaps to determine every one of the traits separately. The information left behind by the people may be utilized to foresee personality, feelings and social emotional traits. However, there are several parts of study carried out in this particular domain, hence as a result we are able to extend and add to this particular study by predicting personality traits by analyzing the reactions to Facebook posts. We also try to extract a bunch of emotional and linguistic features comes up from the reality that owners with various personality traits act psychologically different and make use of various words while getting the point of theirs across

3 Related Works

There have been a number of salient discoveries in the field of personality detection using data collected from a plethora of social media websites. There are various online tools like 'Queendom' and 'Apply Magic Sauce' which gives us an accurate analysis of the user's personality using some predefined questions or social media data. [2] shows us how the personality of a user can be detected using Twitter dataset while [4] and [5] demonstrates personality detection using data scraped from Instagram and Reddit respectively. While the traditional algorithms like k-Nearest Neighbour, k-Means clustering are still in use, [11] shows that the Naive Bayes algorithm gives the best accuracy in predicting an individual's personality using Twitter dataset. [10] talks about a Neural network approach for personality classification. It mainly involves building a Multilayer Perceptron network with five neural networks which acts as a classifier for each of the five personality traits.

4 System Architecture & Algorithms

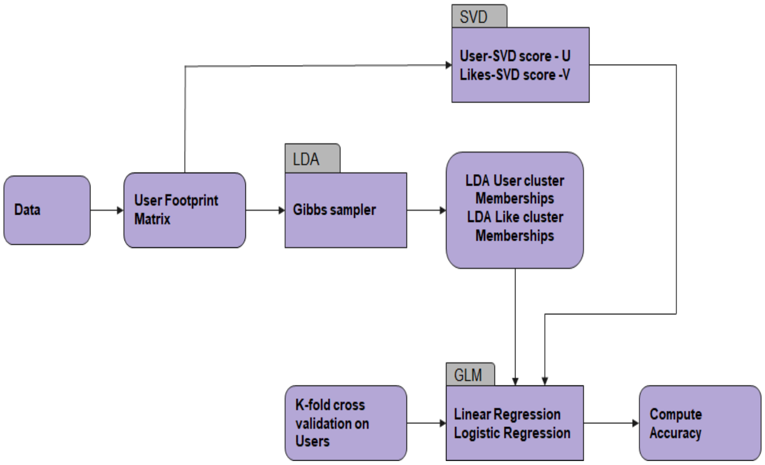

Figure 1. System Architecture

Figure 1 depicts the system architecture of the project. Firstly we perform preprocessing on the dataset, then we will obtain a user-footprint matrix with rows representing users and columns representing likes. This user-footprint matrix is then passed through SVD and LDA dimensionality reduction techniques individually to reduce the number of dimensions and extracted the top 50 dimensions, as we found it is the optimum value since most of the information is retained when the number of dimensions is 50, which will be discussed in later sections. Following this, we will build linear and logistic regression models on the resultant user-footprint matrix with reduced dimensions and apply k-fold cross-validation on the models to avoid overfitting of the data and finally we compute the model's prediction accuracy.

5 Dataset

The dataset has been obtained from the companion website of the “Mining Big Data to Extract Patterns and Predict Real-Life Outcomes" article published in 2016 in psychological methods. This dataset contains profile information and personality traits of Facebook users. The dataset contains:

- **users.csv:** This CSV file consists of the list of 110,728 Facebook users and properties of the users such as a separate ID for each user (userid), age, gender, political view, and big-5 personality trait values.
- **likes.csv:** This file consists of details of 1,580,284 Facebook likes. The details include anonymized IDs(likeid) and names of the likes.
- **users-likes.csv:** This file contains 10,612,326 associations between the users and their likes. It contains two columns, likeid and userid. Each row of this file, which is a likeid and userid pair, implies that a user with a particular userid has liked a page on Facebook with its respective likeid.

6 Pre-processing

The user-like matrix (User Footprint matrix) is constructed in such a way that rows represents users and columns represents likes, and the value one at ith row and jth column of the user footprint matrix says that the ith user has liked jth page in facebook. In our dataset, users have liked a certain number of pages out of the large number of pages present and pages had a definite number of likes given by users out of the large pool of users, so our constructed user footprint matrix turned out to be a sparse matrix. So we have removed the sparse data by keeping users who have liked at least 50 pages and pages which have been liked by at least 150 users. The dimensions of the User Footprint matrix before preprocessing is 110728 x 1580284, and after preprocessing it comes down to 19742 x 8523.

7 Singular Value Decomposition

SVD is a dimensionality reduction technique that is used to express a given matrix (of size m*n) as a product of three matrices:

$$A = U\Sigma V^t$$

Matrix U(m*k) : Left singular vectors
Matrix Σ (square diagonal matrix of size k):
Matrix V(n*k): Right singular vectors

Here, k is the number of dimensions which the researcher chose to extract.

SVD is used to reduce the dimensions of the User Footprint matrix. The input to the SVD block is a user footprint matrix of dimension M and the output is the SVD scores of users and likes for k dimensions.

7.1 Selecting optimum k

There approaches are followed for this[13]:

- Plot the singular values against k. The optimum k is one which lies at the knee of the resulting scree plot.
- Select the k which accounts for 70% of the variance in the original data.

7.2 SVD Rotation

The results obtained from the process of SVD are difficult to understand. To improve the interpretability, SVD dimensions can be rotated. Each of the rotated SVD dimensions contains information which are rich in certain attributes.

To rotate the SVD results, we have used Varimax which is one of the most popular tools for rotation of dimensions. It aims to minimize the number of dimensions related to each of the variables as well as the number of variables related to each dimension.

7.3 SVD Heat Map

Heat maps are plotted to give an idea of the correlations between rotated SVD dimensions and the psychodemo-graphic traits. The traits plotted are the OCEAN traits along with monochotomous variables like political inclination, age and gender.

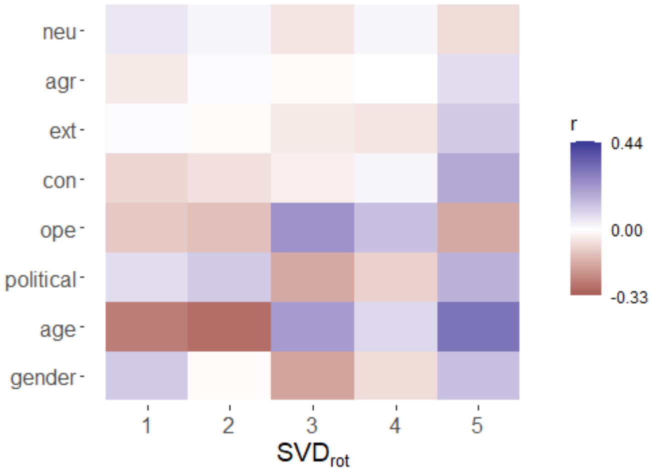

Figure 2: SVD Heat Map showing relation between the rotated SVD dimensions for k=5 and personality traits.

From *Figure 2*, it can be seen that attributes like gender, age and openness are most strongly related to the rotated SVD dimensions. Example: the rotated SVD2 dimension shows a negative correlation with age and the rotated dimension SVD4 depicts a positive correlation with openness and age.

8 Latent Dirichlet Allocation

LDA is one of the most readily interpretable dimensionality reduction techniques, as it produces probabilities unambiguously quantifying the associations between users, footprints, and underlying clusters.[13]

When applied to a matrix of size m*n, it produces matrix γ of size m*k, which describes the probability of each of the users belonging to each of the clusters; a matrix βof size k*n, describing the probability of each of the footprints(personality traits of the user) belonging to each of the clusters.

Like SVD, there are criterias on which the optimum k is selected. The most familiar method is estimating the model's log likelihood estimates. Selecting a k at the end of a rapid log-likelihood value offers decent interpretability of topics. Higher k values provide a better predictive power.

8.1 LDA Heat Map

Similar to SVD, heat maps are plotted to give an idea of the correlations between user's LDA dimensions and user's personality traits. The traits which are plotted here are identical to SVD, namely the OCEAN traits and monochotomous variables like political inclination, age and gender.

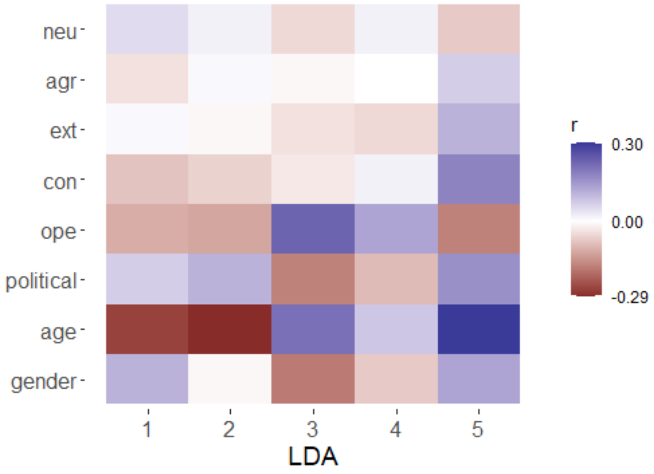

Figure 3: LDA Heat Map to show the correlation between the LDA dimensions and personality traits.

From *Figure 3*, we can see that the dimension LDA1 correlates negatively with age and positively with gender. Dimension LDA2 is the same apart from the fact that it does not correlate with gender. Hence, from the graph we can figure out the relation between personality traits and the LDA dimensions.

9 Evaluations

K-fold validation is applied on the models to prevent overfitting. The data is divided into folds before applying regression models. Each time one of the 10 folds is considered as the test data and the remaining folds as training data. We have used 10-fold validation in this project.

The models used are Linear regression and Logistic regression. To predict the dichotomous variables such as gender, Logistic regression is used and for predicting the multichotomous variables such as personality traits Linear regression is used.

The prediction accuracies are calculated using two evaluation metrics. For dichotomous variables, Area under the ROC curve (AUC) is used and for multichotomous variables Pearson moment correlation coefficient is used.

Variable	**SVD**	**LDA**
Gender (AUC)	.94	.88
Political Views (AUC)	.88	.84
Age	.61	.68
Openness	.44	.42
Conscientiousness	.26	.22
Extroversion	.30	.25
Agreeableness	.24	.18
Neuroticism	.29	.24

10 Visualizations:

Three different analyses have been performed to examine the performance of the models.

- Analyzing accuracy against different dimensions of SVD.
- Analyzing accuracy using SVD and LDA dimensions.
- Analyzing accuracies with different number of users and likes in the user-like matrix.

10.1 Accuracy vs SVD dimensions

Figure 4 shows the line plot of accuracies of all the traits(Age, Agreeableness, Conscientiousness, Extroversion, Gender, Neuroticism, Openness, Political-Factor) against the SVD dimensions(k). We experimented with dimensions ranging from 10 to 150. From the graph, we have observed that when the values of dimensions are between 10 and 50, there is a significant change in the accuracies of all the traits and there is a minute change in the accuracies when the number of dimensions increased beyond 50. So, we have concluded that the optimal number of dimensions for SVD is 50. Moreover, using only the dimensions that retain maximum information helps in increasing the speed of training the model and also produces better accuracies.

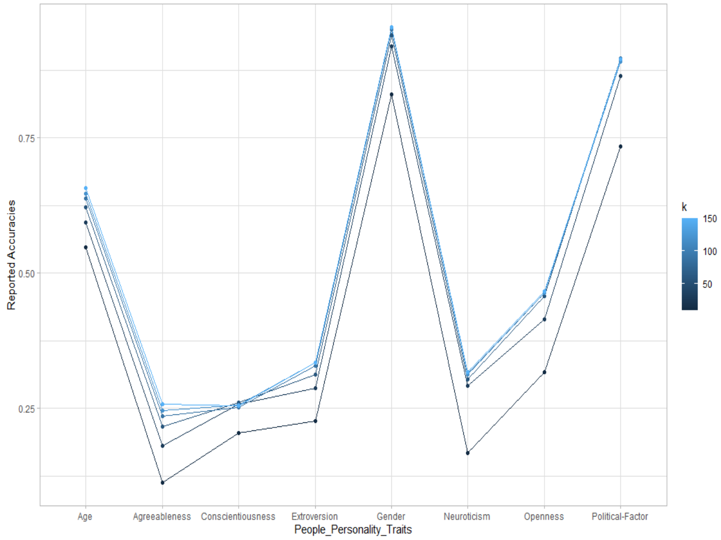

Figure 4 accuracy vs SVD dimensions

10.2 SVD vs LDA:

Figure 5 depicts a line chart comparing the accuracies of models using the dimensions obtained from SVD and LDA for all the traits. We have considered top 50 SVD dimensions/LDA clusters to perform regression. We then observed that SVD performs well when compared to LDA except for the trait 'Age'. Also, the computational time for LDA is very high compared to SVD. So, from this graph we conclude that SVD is the right dimensionality technique for the given scenario/dataset.

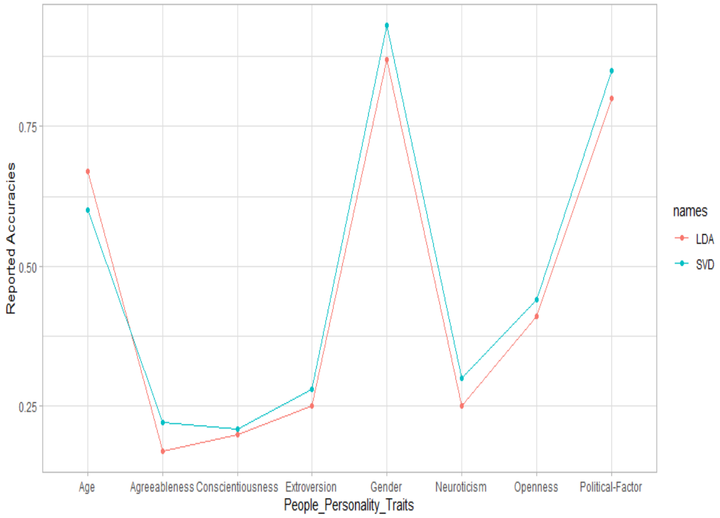

Figure 5 accuracy vs SVD dimensions

11 Variations of user-like matrix:

We have analyzed the accuracies using different threshold values to remove any sparse data of users and facebook likes from the user-like matrix. The threshold values that were found to be optimal were obtained by selecting the users who have liked more than 50 pages and facebook pages which were liked by more than 150 users. If the threshold values were changed beyond this, the change in accuracies were minimal.

12 Division of work and team member's contributions

No	Task	Owner	Start Date	End Date
1	Data collection and analysis	Parimala,Rishitha, Sai Venkat	02/08	02/28
2	Research on clustering techniques and dimensionality reduction	Vineeth, Padma, Nagasai	03/01	03/10
3	Perform dimension reduction	Amit, Sai Venkat	03/11	03/20
4	Apply various clustering algorithms	Parimala, Rishitha, Sai Venkat	03/21	04/02
5	Prepare a model for predicting the personality	Vineeth, Padma, Nagasai	04/03	04/08
6	Analyse the result for classification techniques and draw a conclusion	Amit, Parimala, Rishitha, Vineeth	04/09	04/11

13 Conclusion

After inspecting the results obtained, it can be inferred that for the Big-5 personality traits, the prediction accuracy is comparatively very less.

1. Age and openness which are multichotomous variables possess less prediction accuracies when compared to dichotomous variables like gender and political learning that witnessed higher prediction accuracies due to being less complex.
2. SVD is proven to be more quicker in computing results than LDA when they are configured with the same number of dimensions.
3. The prediction accuracy of a model increases only upto a certain threshold is reached in the number of dimensions used. Thus when more dimensions are added after this, the increase in accuracy is very minimal and negligible.

14 Acknowledgement

The group would like to thank Professor Hasan Davulcu for providing us with an opportunity to work on this project. The project couldn't have been completed without his guidance on each and every step from providing us with the sample dataset to sharing the correct set of resources. Class lectures on dimensionality reduction techniques were extremely helpful in implementing this project. Overall it was a great learning experience for the entire team and we are confident that the concepts learned while implementing this project will help us in our professional carriers.

15 References

[1] https://psycnet.apa.org/fulltext/2016-57141-003.pdf

[2] Golbeck, J., Robles, C. and Turner, K., 2011, May. Predicting personality with social media

[3] Michal Kosinski, Yilun Wang, Himabindu Lakkaraju, and Jure Leskovec, 2016, Vol. 21,

[4] Psychological Methods. Mining Big Data to Extract Patterns and Predict Real-Life Outcomes Using Instagram Picture Features to Predict User's Personality by Bruce Ferwerda, Markus Schedl, and Marko Tkalcic (http://phenicx.upf.edu/system/files/publications/ferwerda_mmm_2016.pdf)

[5] Predicting Character Traits Through Reddit Naval Research Laboratory by Clarissa Scoggins

[6] https://www.pnas.org/content/110/15/5802

[7]https://www.simplypsychology.org/big-five-personality.html

[8] L. C. Lukito, A. Erwin, J. Purnama and W. Danoekoesoemo, "Social media user personality classification using computational linguistic," 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, 2016

[9] Souri, A., Hosseinpour, S. & Rahmani, A.M. Personality classification based on profiles of social networks' users and the five-factor model of personality. Hum. Cent. Comput. Inf. Sci. 8, 24 (2018).

[10] A Neural Network Approach to Personality Prediction based on the Big-Five Model by Dr. Manjula Ramannavar and Dr. Nandini S. Sidnal (https://www.ijirae.com/volumes/Vol2/iss8/09.AUAE10095.pdf)

[11] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," 2015 International Conference on Data and Software Engineering (ICoDSE), Yogyakarta, 2015, pp. 170-174.

[12]https://www.michalkosinski.com/data-mining-tutorial

[13] Mining Big Data to Extract Patterns and Predict Real-Life Outcomes Michal Kosinski, Yilung Wang, Himabindu Lakkaraju, and Jure Leskovec Stanford University