# User Profiling in Social Media

IFT6758 Project - Winter 2019

**Maziar Mohammad-Shahi**
Department of Computer Science and Operational Research
Université de Montréal
Canada
maziar.mohammad-shahi@umontreal.ca

**Maysam Mokarian**
Department of Computer Science and Operational Research
Université de Montréal
Canada
maysam.mokarian@umontreal.ca

**Ryan Mokarian**
Department of Computer Science and Operational Research
Université de Montréal
Canada
ryan.mokarian@umontreal.ca

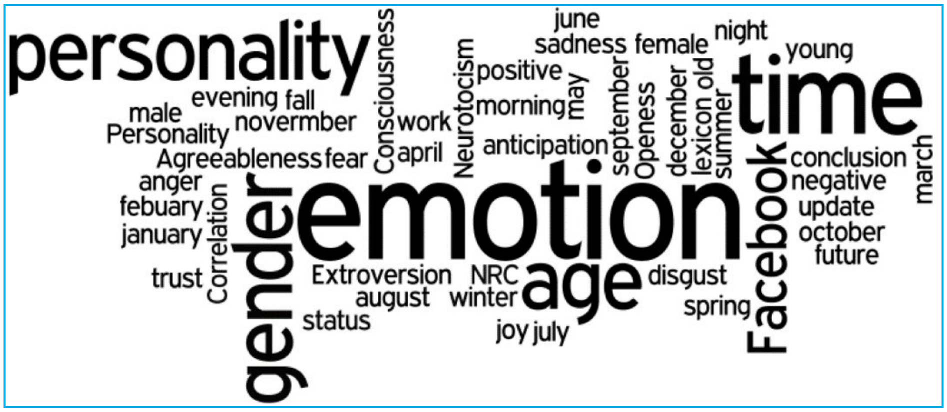


**Figure 1: Word Cloud Diagram***

## ABSTRACT

In this study,[1] we used four different sources of datasets from facebook users to run analysis and perform predictions for various user characteristics. The four datasets included the posts (LIWC, NRC), likes and images(Oxford). The predictions were for age group, gender and the five personality traits of extrovertness, neuroticism, openness, agreeableness and consciousness. In the end, we were able to achieve good results by leveraging the data provided to us. In this report, we will present all of our work and the results that we obtained.

## KEYWORDS

Social Media, Facebook, Personality Traits, Big Five Personality Model, Emotion, Machine Learning, Users' Age, Users' Gender

## 1 INTRODUCTION

Users generate a lot of content in social media such as Facebook. This content, that can be considered as latent variables of social media users, is a rich source of information. It has also acquired a lot of attention as it can be used to identify users' tastes and interests. The users preference can be used for many applications such as in recommendation systems, making policies at government level, companies recruitments, advertising, marketing, sales prediction and so forth. Furthermore, social media includes several sources of information from the users such as text, images, and relations. Usage of multiple sources or modalities of user data helps to arrive at a more accurate user profiles.

[1] This research is a Data Science course (IFT6758) project, presented in Université de Montréal in Winter 2019.

* The figure taken from [1].

In this project, we aim to design a system automatically predict the age, gender, and personality traits value of Facebook users when the users' status, profile picture and "likes" are given as input to the system.

Our results show that image is a good predictor for gender with a relatively high accuracy. For the remaining tasks, we found that a fusion of the various features performs well for both classification and regression tasks.

## 2 METHODOLOGY

This section containsThe machine learning models we used in this research as well as methodologies we used for feature selections.

### 2.1 **Random Forest** (Gender prediction)

This is a tree based machine learning algorithm which works by constructing multiple decision trees. Each tree is created by bootstrapping the data and randomly selecting a subset of features. The prediction is majority vote of decision trees in classification and average in regression tasks.

### 2.2 **Logistic Regression** (Age prediction)

In this machine learning algorithm, we find the coefficients of a linear model. We then apply the logistic function to predict the probability of the output belonging to a class or not.

### 2.3 **Linear Regression** (Personality Trait prediction)

This is a regression method where we try to construct a linear function to model the relationship between the features and the output. As seen in class, we are trying to find the best coefficients to estimate the true relationship.

### 2.4 **Ridge CV** (Personality Trait prediction)

This is a sklearn library which is a linear regression model with L2 regularization with built-in cross validation to facilitate fitting and cross validating the data with fewer method calls. L2 regularization prevents overfitting by penalizing the loss function using an L2 norm of the linear model coefficients.

### 2.5 RFE (Recursive Feature Elimination)

A backward feature selection method that continuously fits and removes the weakest feature until we reach the number of features we want to have in our model.

### 2.6 ICA (Independent Component Analysis)

ICA is a linear dimension reduction method to detect latent variables of data. For feature reduction, it decomposes multivariate signals of data into independent non-Gaussian signals (dimensionally reduced independent features).

## 3 DATASET AND METRICS

This section contains the datasets we are using in this research as well as analysis on each dataset..

### 3.1 NRC[2] and Users Personality Profile[3]

#### *3.1.1 Data frame preparation and checking for missing data.*

Information of 9500 Facebook users from nrc.csv and Profile.csv are used in this section. nrc.csv file includes the following 10 emotional features which are considered as predictors.

- Positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, trust.

Profile.csv includes the following 7 predicates.

- Age, gender, ope, con, ext, agr, neu.

The last 5 represent the big five personality traits: Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism (Emotional Stability) [5].

For pre-processing, first datasets from both Profile.csv and nrc.csv files are merged on the

[2] NRC is a lexicon that contains more than 14,000 distinct English words annotated with 8 emotions (anger, fear, anticipation, trust, surprise, sadness, joy,
and disgust), and 2 sentiments (negative, positive). For more information, refer to [2].

[3] Users Personality Profile are five traits of the Big Five personality model, namely Openness to experience, Conscientiousness, Extroversion, Agreeableness, and Emotional Stability. They are scored between one to five. For more information, refer to http://mypersonality.org and [5].

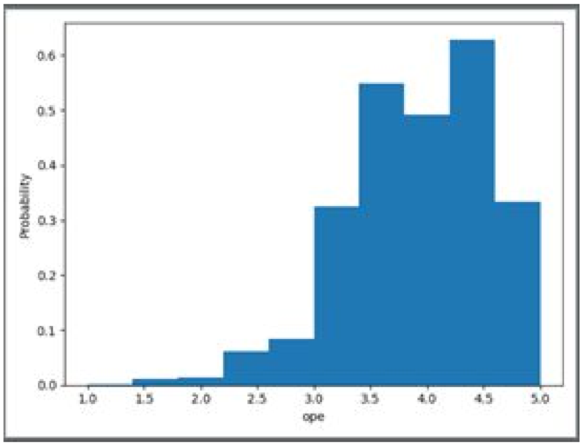
userid. Then, we check if there is any missing value. The results, summarized below, shown that there is no missing values on the datasets:

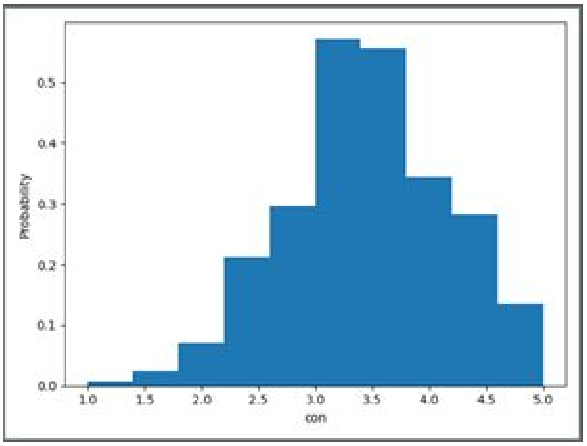
- Length of the original data frame: 9500
- Length of the data frame after removing missing values: 9500
- Number of rows with at least 1 NA value: 0

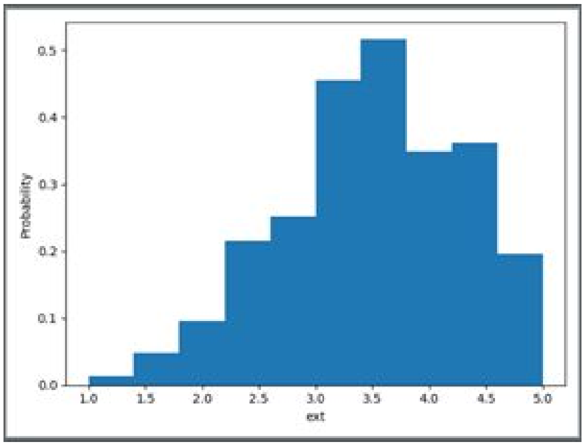
There were however some outliers. We saw that there were 21 users over the age of 100 which are most likely incorrectly entered ages and can potentially reduce our accuracy for age prediction.

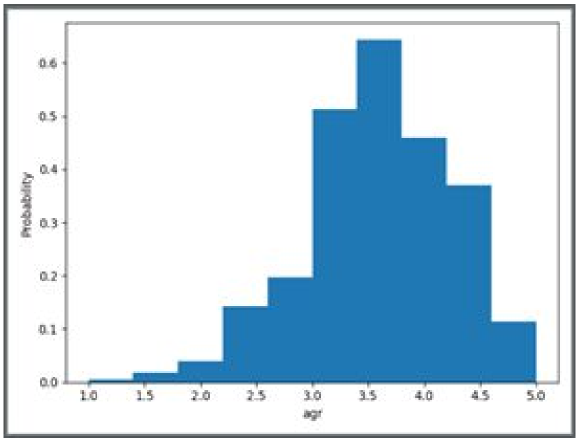
### *3.1.2 Personality Traits Probability Distribution.*

Probability distribution of personality traits are shown Fig. 2.









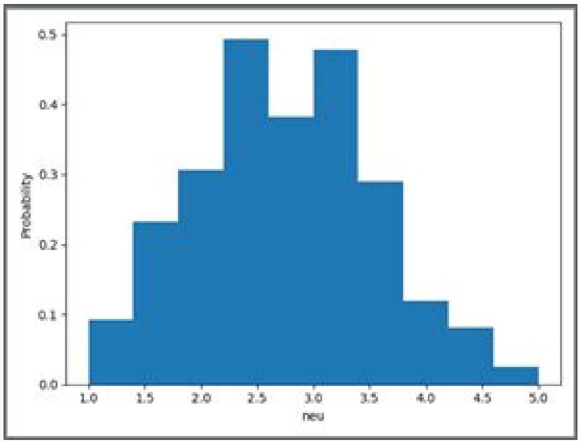


**Figure 2: Probability Distribution of Personality Traits**

To establish a baseline, Mean and SD of all users' scores for emotion features and dependent variables are prepared and presented in the following sections.

### *3.1.3 Personality Traits Mean and Standard Deviation (SD).*

**Table 1: Age, gender and personality traits Mean and SD**

| | age | gender | ope | con | ext | agr | neu |
|---|---|---|---|---|---|---|---|
| Mean | 26.42 | 5483 female | 3.91 | 3.45 | 3.49 | 3.58 | 2.73 |
| SD | 10.38 | 4017 male | 0.63 | 0.72 | 0.81 | 0.66 | 0.79 |

Mean of the personality traits and emotion features are shown on the below radar chart.

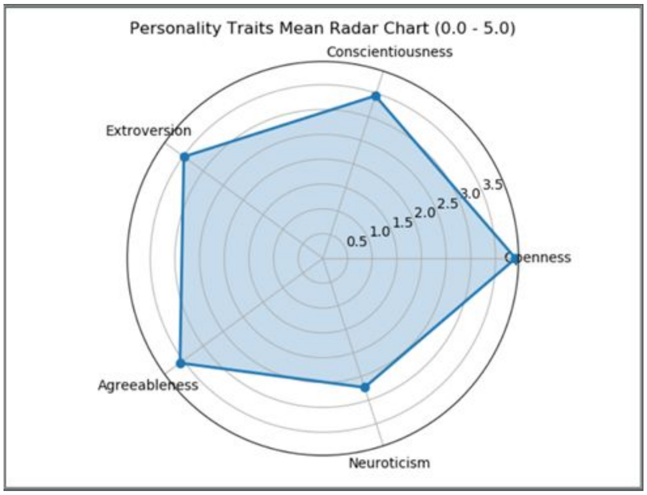


**Figure 3: Personality Traits Mean Radar Chart**

### *3.1.4 NRC Emotion Features Probability Distribution.*

Probability distribution of emotion features are shown in Fig. 4 where it has been processed by log+1 and Min Max scaling. It is observed that there is skewness toward left for emotions with negative stigmas such as anger, fear, disgust, meaning most people presented low scores in these features. Other features more or less follow a normal curve distribution where the means of the distributions have a shift to left. Exception is the “positive” feature which its mean has a shift to the right, meaning on average people presented high score for this feature.

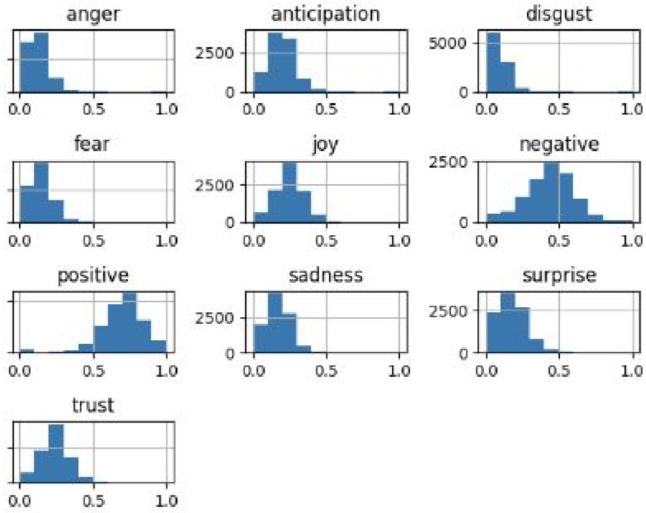


**Figure 4: Probability Distribution of Emotion Features**

### *3.1.5 NRC Emotion Attributes Mean and Standard Deviation (SD).*

**Table 2: Emotion attributes Mean and SD**

| | Pos | Neg | Ang | Ant | Dis |
|---|---|---|---|---|---|
| Mean | 0.63 | 0.36 | 0.09 | 0.15 | 0.06 |
| SD | 0.17 | 0.16 | 0.07 | 0.08 | 0.06 |

| | Fea | Joy | Sad | Sur | Tru |
|---|---|---|---|---|---|
| Mean | 0.11 | 0.19 | 0.13 | 0.07 | 0.18 |
| SD | 0.07 | 0.09 | 0.07 | 0.05 | 0.09 |

Note: Pos: Positive, Neg: negative, Ang: anger, Ant: anticipation, Dis: disgust, Fea: fear, Sad: sadness, Sur: surprise, Tru: trust

Fig. 5 illustrates a radar chart for the emotion features.

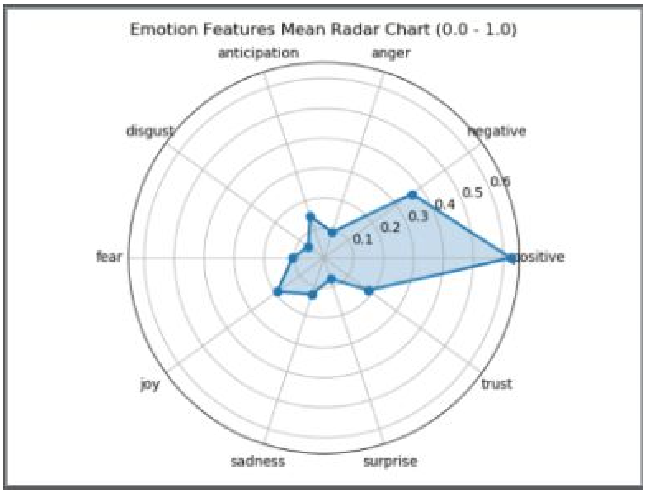
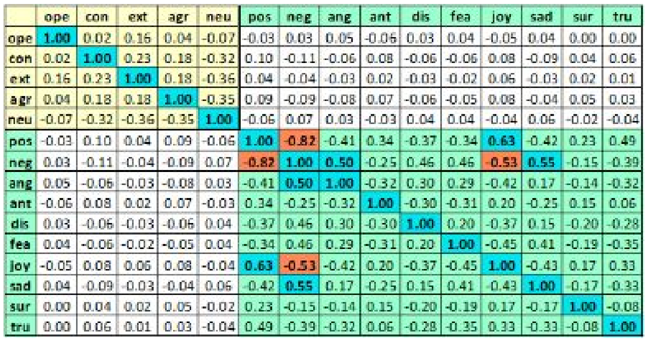


**Figure 5: Emotion Features Mean Radar Chart**

*3.1.6 Pearson Correlation results between each pair of Emotion Features & Personality Traits.*

Table 3 presents the Pearson Correlation between the different personality traits and emotion features. As it is shown, there is not a high correlation between each personality trait and each emotion feature. Among emotion features, we observe a high negative correlation between positive and negative, which makes sense as they are opposite emotions. Other interesting findings are positive correlation between "positive" and "joy" as well as between "negative" and "sadness", which again they are trivial.

**Table 3:Personality Traits - Emotion Features Correlation Matrix**



| | ope | con | ext | agr | neu | pos | neg | ang | ant | dis | fea | joy | sad | sur | tru |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ope | 1.00 | 0.02 | 0.16 | 0.04 | -0.07 | -0.03 | 0.03 | 0.05 | -0.06 | 0.03 | 0.04 | -0.05 | 0.04 | 0.00 | 0.00 |
| con | 0.02 | 1.00 | 0.23 | 0.18 | -0.32 | 0.10 | -0.11 | -0.06 | 0.08 | -0.06 | -0.06 | 0.08 | -0.09 | 0.04 | 0.06 |
| ext | 0.16 | 0.23 | 1.00 | 0.18 | -0.36 | 0.04 | -0.04 | -0.03 | 0.02 | -0.03 | -0.02 | 0.06 | -0.03 | 0.02 | 0.01 |
| agr | 0.04 | 0.18 | 0.18 | 1.00 | -0.35 | 0.09 | -0.09 | -0.08 | 0.07 | -0.06 | -0.05 | 0.08 | -0.04 | 0.05 | 0.03 |
| neu | -0.07 | -0.32 | -0.36 | -0.35 | 1.00 | -0.06 | 0.07 | 0.03 | -0.03 | 0.04 | 0.04 | -0.04 | 0.06 | -0.02 | -0.04 |
| pos | -0.03 | 0.10 | 0.04 | 0.09 | -0.06 | 1.00 | -0.82 | -0.41 | 0.34 | -0.37 | -0.34 | 0.63 | -0.42 | 0.23 | 0.49 |
| neg | 0.03 | -0.11 | -0.04 | -0.09 | 0.07 | -0.82 | 1.00 | 0.50 | -0.25 | 0.46 | 0.46 | -0.53 | 0.55 | -0.15 | -0.39 |
| ang | 0.05 | -0.06 | -0.03 | -0.08 | 0.03 | -0.41 | 0.50 | 1.00 | -0.32 | 0.30 | 0.29 | -0.42 | 0.17 | -0.14 | -0.32 |
| ant | -0.06 | 0.08 | 0.02 | 0.07 | -0.03 | 0.34 | -0.25 | -0.32 | 1.00 | -0.30 | -0.31 | 0.20 | -0.25 | 0.15 | 0.06 |
| dis | 0.03 | -0.06 | -0.03 | -0.06 | 0.04 | -0.37 | 0.46 | 0.30 | -0.30 | 1.00 | 0.20 | -0.37 | 0.15 | -0.20 | -0.28 |
| fea | 0.04 | -0.06 | -0.02 | -0.05 | 0.04 | -0.34 | 0.46 | 0.29 | -0.31 | 0.20 | 1.00 | -0.45 | 0.41 | -0.19 | -0.35 |
| joy | -0.05 | 0.08 | 0.06 | 0.08 | -0.04 | 0.63 | -0.53 | -0.42 | 0.20 | -0.37 | -0.45 | 1.00 | -0.43 | 0.17 | 0.33 |
| sad | 0.04 | -0.09 | -0.03 | -0.04 | 0.06 | -0.42 | 0.55 | 0.17 | -0.25 | 0.15 | 0.41 | -0.43 | 1.00 | -0.17 | -0.33 |
| sur | 0.00 | 0.04 | 0.02 | 0.05 | -0.02 | 0.23 | -0.15 | -0.14 | 0.15 | -0.20 | -0.19 | 0.17 | -0.17 | 1.00 | -0.08 |
| tru | 0.00 | 0.06 | 0.01 | 0.03 | -0.04 | 0.49 | -0.39 | -0.32 | 0.06 | -0.28 | -0.35 | 0.33 | -0.33 | -0.08 | 1.00 |

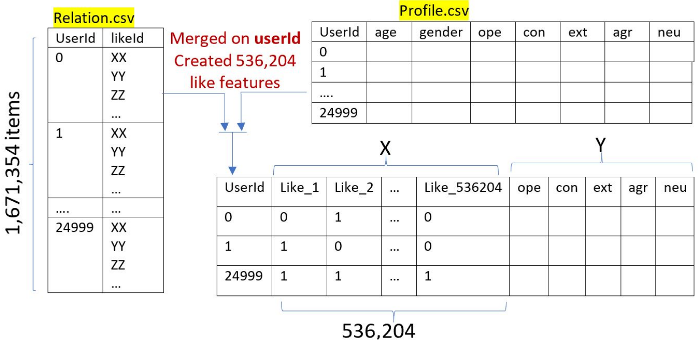
Note contractions of titles on the table are described below:

Open (ope), Conscientious (con), Extrovert (ext), Agreeable (agr) and Neurotic (neu)

positive (Pos), negative (Neg), anger (Ang), anticipation (Ant), disgust (Dis), fear (Fea), joy (Joy), sadness (Sad), surprise (Sur), and trust (Tru)

### 3.2 Facebook Page Likes[4]

Facebook users' Page likes is our relational dataset modality. We try to use it for predicting users' personality traits. The data was given in Relations.csv file with 1,671,354 rows (user's like items). Relations.csv file is merged with Profile.csv file (included users five personality traits, age and gender) on user ids. As illustrated on Fig. 6, using a one-hot encoding method we converted the merged data to a matrix where rows represent user ids and columns represent like pages (as independent variables) and personality traits (as dependent variable). There are 25000 users with 536,204 unique page likes. Value of each matrix entry is one if the user (row) likes a respected page (column) in the dataset, otherwise it is zero.



| UserId | likeId |
|---|---|
| 0 | XX<br>YY<br>ZZ<br>… |
| 1 | XX<br>YY<br>ZZ<br>… |
| …. | … |
| 24999 | XX<br>YY<br>ZZ<br>… |

| UserId | age | gender | ope | con | ext | agr | neu |
|---|---|---|---|---|---|---|---|
| 0 | | | | | | | |
| 1 | | | | | | | |
| …. | | | | | | | |
| 24999 | | | | | | | |

| UserId | Like_1 | Like_2 | … | Like_536204 | ope | con | ext | agr | neu |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | … | 0 | | | | | |
| 1 | 1 | 0 | … | 0 | | | | | |
| 24999 | 1 | 1 | … | 1 | | | | | |

**Figure 6: Users "Page Like" and "Profile" Merge and Conversion to one-hot encoding matrix**

### 3.3 LIWC[5]

For this dataset, we found that there are some features that can be potentially excluded for age prediction because they do not distinguish between age groups or have very little effect and are therefore

[4] Users prediction can be done by indirect relations among users, i.e. shared Facebook page likes. For more information on the "Page Likes" model, refer to [6].

[5] LIWC (Linguistic Inquiry and Word Count). It is a well-known standard in computerized text analysis which includes features related to standard counts, psychological processes, relativity, personal concerns, and linguistic dimensions. For a complete overview of the LIWC features, refer to [3].

not useful. This was commonly found in a lot of the ponctuation-related features in the LIWC dataset. Figure 7 and 8 presents two of such cases.

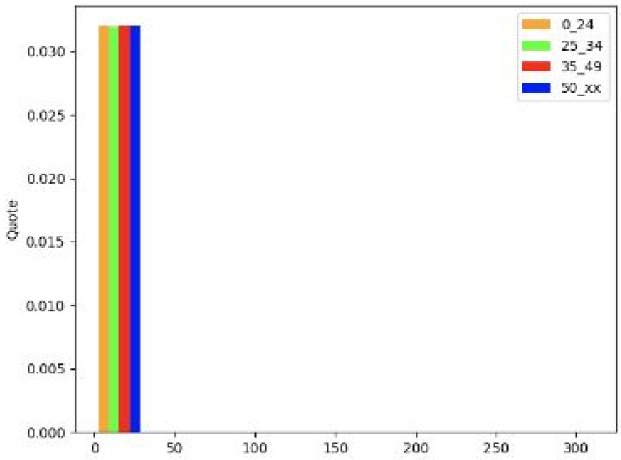


**Figure 7: Age Group Histogram with respect to Quote**

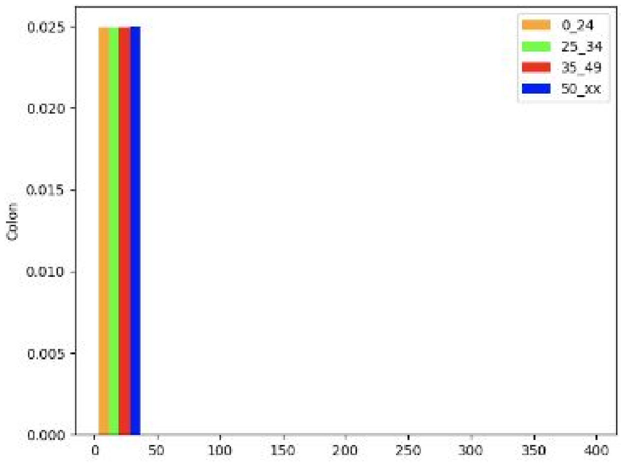


**Figure 8: Age Group Histogram with respect to Colon**

No nulls were found in the data and the types did not need fixing since after reading the data, the data types were all numbers as expected.

For preprocessing, we noticed that the LIWC features are not all on the same scale. We first applied a min/max normalization on the data.

There were many cases of skewness similar to what we see below in figure 9 in the data. To improve this, we also applied a log plus one transformation on the dataset.

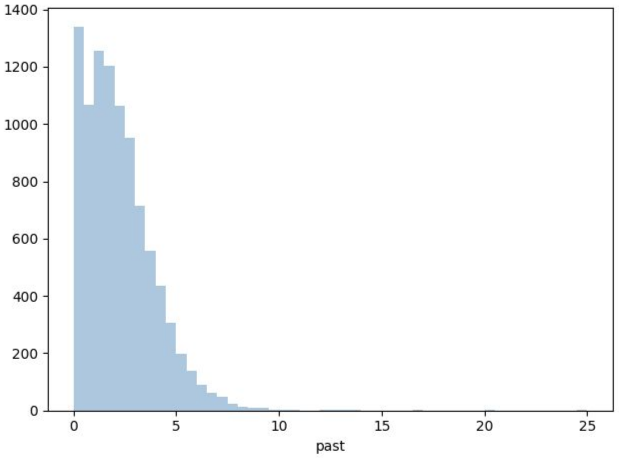


**Figure 9: Skewness in LIWC Features**

As it will be shown in the next sections, to get rid of the features that are not good predictors, we used recursive feature elimination (RFE).

### 3.4 Oxford[6]

This dataset consists of 64 facial features. We analyzed the facial components by their interaction with age, gender and personality traits. We splitted the facial features into: facial_recatngle_*, pupul_left_*, nose_tip_*, mouth_*. eyebrow_*, eye_*, *_lip_* and facialHair_*. We noticed that facialHair_* features such as facialHair_beard, facialHair_sideburns and facialHair_mustache are highly interactive with gender prediction.

figure 10, (top plot) indicates the facialHair_* feature interaction with age, gender and personality traits. Additionally The bar plot (middle and bottom of the figure), illustrates the facialHair features of the male (green) and female (orange). Obviously there is a significant difference in the gender interaction and facialHair_* features and these data is being used to train our model in section 4.1

[6] Oxford includes facial features such as face rectangle features, face landmark features, face characteristics including age, gender, facial hair, smile, head position and glasses type. For more information, refer to [4].

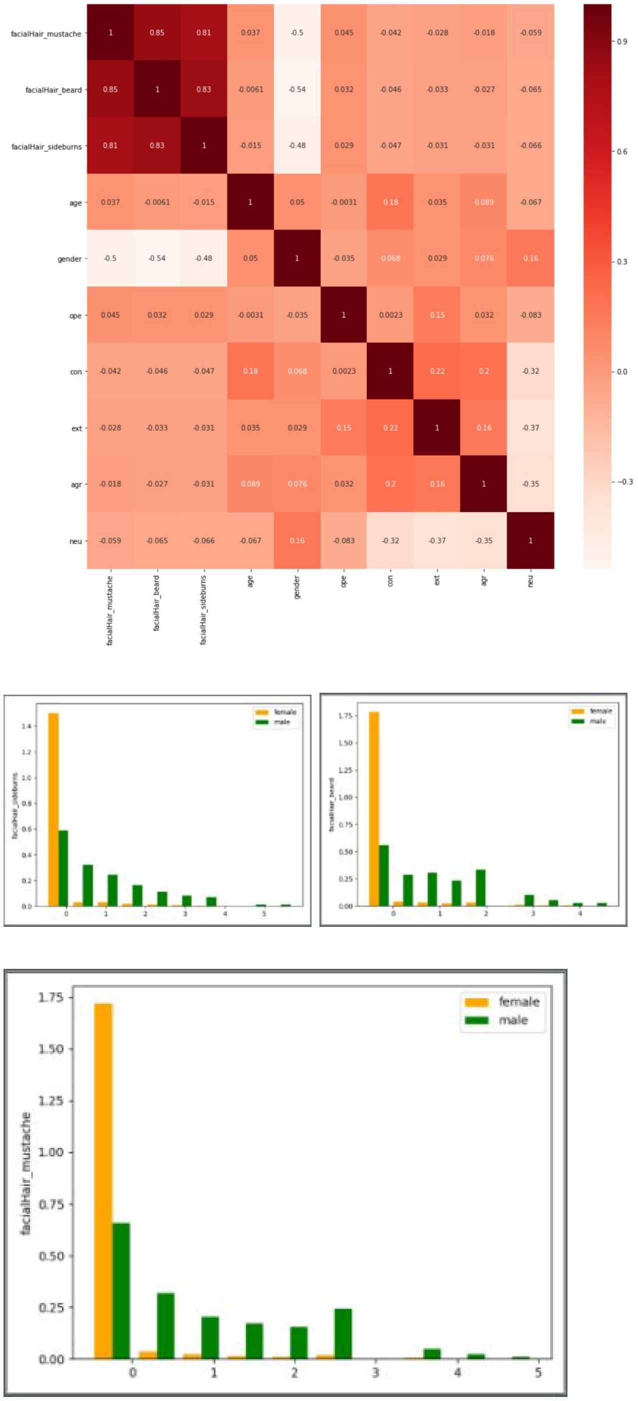


**Figure 10: (plot on top)FacialHair_* interaction with age, gender and personality traits. (plot on middle/bottom) bar plots of female/male distribution over different facialHair_* features**

## 4 RESULTS

After performing analysis on the different datasets and applying feature engineering and feature selection, we executed the experiments to predict age, gender and personality traits. This section consists of the results for several machine learning experiments and the tuned hyperparameters on predicting age, gender and personality traits.

### 4.1 Gender

For the gender prediction task, we tried multiple different models. In the end, we selected to use random forest as it produced the best results on the image features. We initially got around 78% accuracy on the validation data. We then performed a grid search on the hyperparameters of max depth and number of trees (n_estimators) and got to increase the accuracy to 84%. The results of our experiments are presented in Table 4 below.

**Table 4: Gender Accuracy Result**

| Feature | Hyper parameter | Model | Accuracy |
|---|---|---|---|
| Images | max_depth=10<br>n_estimators=400 | Random Forest | **0.84** |
| Images | loss="hinge"<br>penalty=12 | SGD | 0.62 |
| Images | gamma=1.0 | Radial Basis (RBF Kernel) | 0.59 |
| Images | k=3 | KNN | 0.58 |
| Likes | k=6 | KNN | 0.58 |

### 4.2 Age

Summary of Age accuracy results shown on Table 5. For datasets, a combination of LIWC and Oxford was used. The shown accuracy is for the Logistic Regression model with different amounts of features. The best accuracy found for 130 features.

For the feature selection, Recursive Feature Elimination (RFE) and Independent Component Analysis (ICA) were used. As mentioned earlier, since there are a few features that are not good predictors of age, RFE helped increase the accuracy by getting rid of those features.

A hyperparameter optimization using grid search was also performed and resulted to the below values that were used in the final model:

- Penalty='l2'
- C=0.00483
- Max_iter=100

Note a 10 Fold splits Cross Validation was used.

**Table 5: Age Accuracy Result**

| Model Features | Accuracy |
|---|---|
| All features | 61.84% |
| 10 features where ICA applied | 59.65% |
| 130 features where RFE applied | **62.4%** |

### 4.3 Personality Traits

Summary of Personality Traits error rates shown on Table 6. We tested several models with different datasets. The table shows the final models and dataset that gave us the lowest error rates. Note we used Log Transformation and Min-Max scaling for data Pre-Processing and a 10 Fold splits Cross Validation for data Post-Processing. The data column presents the data used in the model to obtain the corresponding error rate.

**Table 6: Personality Traits Sqrt of MSE**

| Model | Data | Ope | Con | Ext | Agr | Emo Stb |
|---|---|---|---|---|---|---|
| LinearRegression | LIWC,NRC | **0.629** | **0.71** | 0.816 | **0.662** | 0.790 |
| LinearRegression | LIWC,NRC,Oxford | - | - | **0.815** | - | - |
| RidgeCV | LIWC,NRC,Oxford | - | - | - | - | **0.783** |

Note: Ope: Openness, Con: Conscientiousness, Ext: Extroversion, Agr: Agreeableness, Emo Stb: Emotional Stability

## 5 CONCLUSION AND FUTURE WORK

In summary, we were able to get good accuracy on the age and gender prediction tasks and beat the baselines. We were also able to mostly decrease the baseline error rate of the personality traits. One thing that seems promising was taking better advantage of the likes features to further improve our models. It seems that we were running into dimensionality issues with the likes and given more time, we would invest that into reducing the dimension to mitigate that problem. Another opportunity, with which we briefly experimented, is using word2vec for the likes to embed them and use that as a similarity measure for users. Given more time, we believe that we would have been able to improve our models and predictions for all the tasks.

## REFERENCES

[1] G. Farnadi, G. Sitaraman, M. Rohani, M. Kosinski, D. Stillwell, M. F. Moens, S. Davalos, and M. De Cock. 2014. How are you doing? Emotions and Personality in Facebook. *EMPIRE 2014 workshop-2nd Workshop on "Emotions and Personality in Personalized Services.* https://empire2014.files.wordpress.com/2014/08/02_empire_2014_farnadi.pdf

[2] S. Mohammad, X. Zhu, J. Martin. 2014. Semantic role labeling of emotions in tweets. In: Proceedings of the WASSA, pp. 32–41.

[3] Y.R. Tausczik, J.W. Pennebaker. 2010. The Psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. 29, 24–54.

[4] Z. Cao, Q. Yin, X. Tang, and J. Sun. 2010. Face recognition with learning-based descriptor. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2707–2714. IEEE.

[5] P. T. Costa, R. R. McCrae. 2008. The revised NEO personality inventory (NEO-PI-R). The SAGE Handbook Of Personality Theory And Assessment, 2:179– 198.

[6] M. Kosinski, D. J. Stillwell, T. Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. Proc. of the National Academy of Sciences (PNAS), 110:5802–5805.