SOCIAL MEDIA USER PROFILING

JOSS RAKOTOBE, Université de Montréal
COLIN NDONFACK, Université de Montréal
WRUSHABH WARSHE, Université de Montréal

ACM Reference Format:
Joss Rakotobe, Colin ndonfack, and Wrushabh Warshe. 2018. SOCIAL MEDIA USER PROFILING. *ACM Trans. Graph.* 37, 4, Article 111 (August 2018), 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In this report we aim at building an system for automatic recognition of the age, gender, and personality of social media user. The system need to “predict” the following information about the user as output:

- GENDER: either “male” or “female”
- AGE: either “xx-24”, “25-34”, “35-49”, or “50-xx”
- PERSONALITY: score between [1,5] for each of the five traits of the Big Five personality model, namely Openness to experience, Conscientiousness, Extroversion, Agreeableness, and Emotional Stability (reversely referred to as Neuroticism).

Predicting gender is a binary classification or concept learning task. Predicting age is a multi-classification task.Whereas, predicting personality is a regression task. We succeeded in achieveing 87% accuracy on the gender prediction and 65.6% for the age prediciton task. Moreover we obtained RMSE of 0.651 for Openness to experience, 0.794 for Neuroticism, 0.783 for Extroversion, 0.657 for Agreeableness, 0.722 for Conscientiousness personality traits.

The general flow of this report is as follows: The dataset and metrics section briefly talks about the data provided . There are 3 separate section for each classification task namely: gender classification task, age classification task and personality classification task. Each of these section have a methodology sub-section, a feature engineering sub-section and a result sub-section. Methodology sub section will talk about the the Machine learning models used. The feature engineering sub-section will talk about the data preprocessing, feature extraction, feature selection methods. And, the result sub-section will discuss about all the various methods and model attempted before reaching to the final/best model. It will also discuss why some methods failed and why some worked. Additionally we will try to address some of the issues we faced why implementing some models. The conclusion section will briefly summarize the final result and discuss the future methods which could be interesting to try.

2 DATASET AND METRICS

The dataset contains personal information about Facebook users. They were divided into three parts: the profile picture of the users, their messages and their liked pages.

- The profile pictures were already pre-processed into Oxford features for privacy issue. These are features represents some facial points of the face present in the profile pictures (see figure 1). In addition to the facial points we also have the angle of rotation of the face, some facial hair features and the size of the picture. In total we have 64 features.

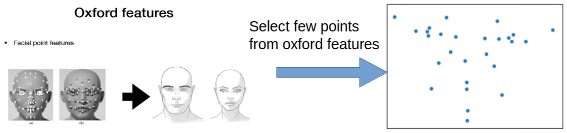

Fig. 1. Oxford feature points of the profile pictures

- The messages were also preprocessed into Linguistic Inquiry and Word Count (LIWC) and NRC features, again for privacy issue. LIWC features is a large set of features describing the number of words, the tense, grammatical type of word, punctuation, emoticons used, and more.
  NRC features are more focused on sentiments : positive, negative, joy, fear, etc. In total, we have a set of 91 features
- The likes is a dataset containing the page ID of the pages each users liked.
  For our prediction, we used two type of metrics : the accuracy for the age and the gender classification, and the root-mean-square error (RMSE) for the regression on the personalities.
  The accuracy is the ratio of well-classified examples over the total number of examples.
  Since it is almost impossible to predict a continuous value, we use prefer the RMSE for regression tasks. The RMSE is the square root of the mean of the squared errors between each prediction and its real value. Unlike the accuracy, the lesser its value, the better the model.

3 GENDER CLASSIFICATION TASK

3.1 Methodology

For gender classification we applied Naive bayes model on the users having profile picture and the MLP model on users with missing profile pictures. To predict the classes for users with no profile picture we used the pages liked data. Therefore, we trained the MLP model on the relation data (pages liked by each user). The final accuracy for the gender classification task was then reported as a weighted mean for the both the model: Naive bayes and MLP.

Authors’ addresses: Joss Rakotobe, Université de Montréal; Colin ndonfack, Université de Montréal; Wrushabh Warshe, Université de Montréal.

0730-0301/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

3.2 Feature Engineering

Recall, in this task we want to predict the gender of the user which can be either a male or a female.

- Male - Class 0,
- Female - Class 1,

In order to predict the gender we decided to rely on the profile pictures and the pages liked by the user. The profile picture information comes in the form of "oxford.csv" data format which gives us the set of key facial points on a users face.The liked pages information comes from the relation data which contains page id of all the pages a user has liked.

Even before performing feature engineering on the dataset we need to clean the data first.

- The pre-processing steps in order to get clean data is mentioned below:
  - Remove the irrelevant features: headPose_pitch (because all values=0).
  - Remove duplicate features (FaceRectangle_height = FaceRectangle_width).
  - If more than 1 profile image appears for the same single user, select the largest image. This selection originated from our own observation and intuitive understanding. The reasoning behind this is that in general profile pictures will have the user face covering larger portion of the picture and other entities (if present) will occupy smaller area in the profile picture. Hence if multiple entries of the profile picture are present in the dataset we selected the entry which gave us the largest profile picture. Of course this could go wrong in some select cases but for majority it seems to work well.
  - For the facial hair feature for female gender we set its value to zeros. This comes from our observation that the female gender generally do not have facial hair.

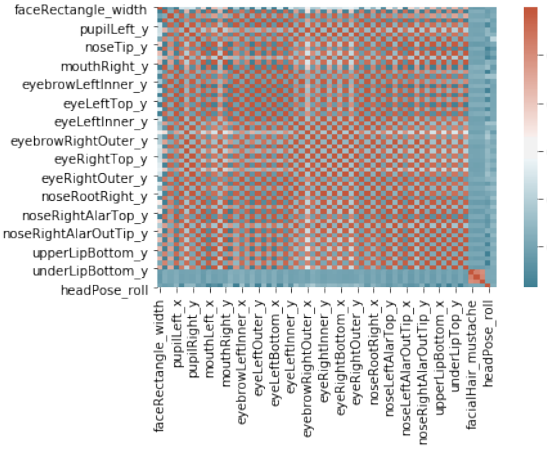

Fig. 2. Correlation matrix on all the original feature.(The figure shows a 64*64 matrix but the labels are only given for 16 feature for better readability.)

- To check if the data is redundant, in other words, are the features highly related to each other, we plot a correlation heat map over all the original feature as shown in the figure 2.
  - From the figure 2 one can see that most of the features are highly correlated.
  - In the figure 3, which is a subsection of figure 2, we can see the features which are not correlated. We could just use these 6 features for classification task but apparently they are not discriminative feature.You can see in the figure 4. The box plot distribution (mean, deviation, spread) of the feature headpose_yaw is same for both the gender, indicating that the features are not at all discriminative. Hence we need to perform feature engineering and obtain more non-redundant feature set.

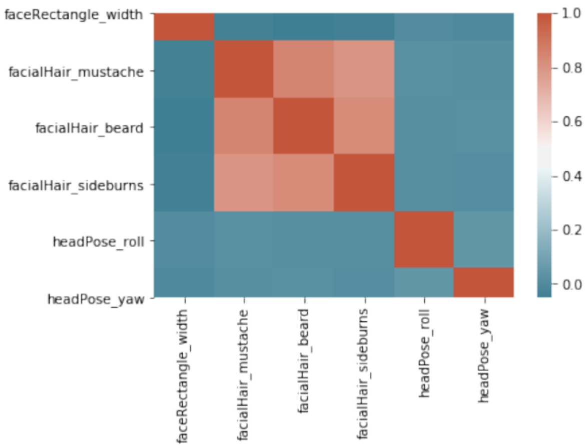

Fig. 3. Correlation map of the select feature from the above figure 2 bottom right corner

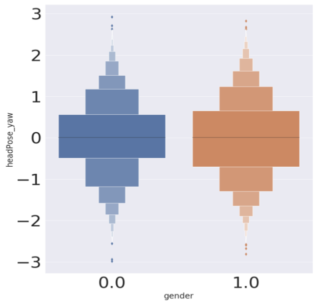

Fig. 4. Box plot of the Headpose yaw feature. (Orange color for female gender and blue for male gender)

- Now we have more cleaner data than before. Moving forward, instead of using all the oxford feature points(since they are redundant and non-discriminative as seen above) we decided to extract some rich feature. We began by choosing key points in the profile picture. Key points are those points which provides significant information about the face structure. As shown in the figure 5 the key points are located on the anticipated parts of a human face. For example the corner of lips, the corner of eyes, the corner of eyebrows, the corner of nose etc.

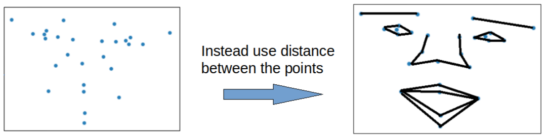

Fig. 5. Selection of feature points from the original oxford feature points.

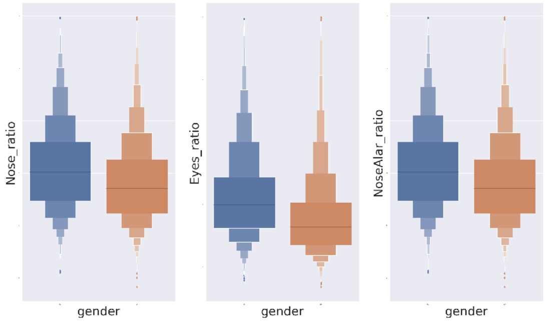

Fig. 6. Boxplot of the select ratio feature to check if they are discriminative (Orange color is for female gender and the blue color is for male gender)

- Moreover, after some experiment with using just the key points we realized that using distance between the key points could be more beneficial as they could be more discriminative in nature. And this proved to be the right assumption because we observed the following points in the data
  - Eyebrow width and height tends to be higher for female than for the male counterpart.
  - Nose height and nose width tends to be smaller for female gender.
  - Lips width and lips height tends to be larger for Female gender.

Once we have the distances between the key points we additionally took the ratio between the relevant feature distance and obtained a new set of features. The ratio were selected on semantic information. Some of the ratios are given below:

- Ratio of lips width to lips height
- Ratio of nose width to nose height
- Ratio of left eye width to left eye height. Same for the right eye.

This new set of features turned out to be the most rich and discriminative set of features.

In all, our final set of features contains the distance between the key points and ratio of these distance feature. To verify if all the above feature selection and feature extraction method actually gave us a set of discriminative feature we use box-plot as shown in figure 6. Looking at the distribution (mean, variance, spread) of the box-plot for various features we can confidently say that the chosen ratio feature are indeed discriminative.

3.3 Result

The models we finally chose to apply for gender classification task is a combination of Naive Bayes model and Multi player perceptron (MLP) model. The reason to choose these individual models are as follows and the reason to choose the combination will be explained in the following paragraph:

- Naive Bayes model (NB): In order to visualize how the features are distributed we randomly selected few features and observe the scatter plot. r. We observed that the features are distributed normally (Gaussian), see figure 7. Naive Bayes model works on the assumption that the features are distributed normally. Hence, we proceeded to chose the Naive Bayes model. Note that all the feature except for the feature facial hair follows Gaussian distribution. Due to its non normal distribution, the feature facial hair turned out to be the discriminative feature and helps the model in classification.

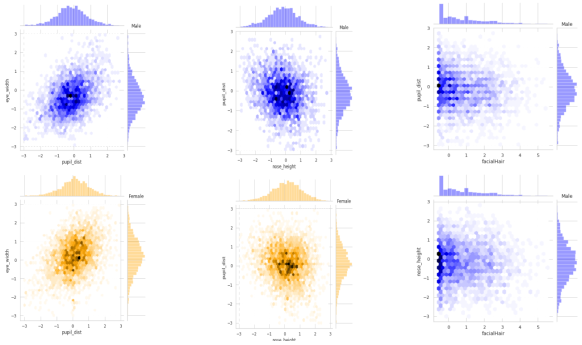

Fig. 7. Distribution of the features (On the x and y axis are the features. Blue color for male gender and orange color for female gender. Selected features are eye width, pupil dist, nose height and facial hair)

- Multi-layer preceptron (MLP): Some of the entries (users) in the profile picture data were without any kind of profile picture information. In order to deal with those entries we decided to utilize another dataset which is pages liked (relation dataset). And the relation dataset features were not distributed normally and was destined to give poor result if Naive bayes model were supposed to be used. Hence, we decided to use MLP which can give us non-linear mapping from the dataset to the label class of the relation data. We trained our MLP model on the pages liked dataset.

Please refer to the Table 1 for the final results. Keep in mind that the accuracy of around 88% ,see table 1, using only the NB model is actually an average accuracy over the complete dataset. Which is basically a weighted average of very high accuracy on user with profile pictures and very low accuracy on users without profile picture. And the accuracy has higher value just because most of the users has profile pictures. We wanted to build a system which does not give poor accuracy for the user without profile picture. Therefore, we decide to go for the combination of 2 models : NB and MLP. Combination model reports the weighted accuracy (87%) by using NB on users with profile picture and using MLP on the users without profile picture. In this case the accuracy on the users without profile picture is comparatively very high. And the overall accuracy is around 87%. Such model gives best of both the world and stays fair to all kinds of user.

Note that in the one of the pre-processing steps we assigned zero to the facial hair feature for the female gender. Such manipulation is actually an algorithmic bias. Bias is against the female gender with the facial hair. So we can say that the model is discriminative in nature in some sense. Even though such pre-processing step improves the over all accuracy we should keep the discriminative bias as minimum as possible. It would be interesting to see how much such bias affects the performance in the future work.

4 AGE CLASSIFICATION TASK

4.1 Methodology

The task was to predict - given user photos, page likes, text messages and emoji; at least some already extracted data - which age group does the user belong to. Let's recall that the group considered here are :

- up to 24 years old (xx-24),
- from 25 to 34 years old (25-34),
- from 35 to 49 years old (35-49) and
- at least 50 years old (50-xx).

We used only relation data set and the model applied was the neural network, especially Multi-Layer Perception (MLP) classifier. As main parameters of this MLP model, we used, the rectified linear unit (ReLU) activation function, and hidden layers of one hundred neurons. As it is a multi-label classification task, the output function is obviously the softmax function. The choice of this model relied on the fact that it is more a convenient method when we have too many features comparing to sample size. In our case, we had more than 500000 one hot encodings of page likes for just 8500 individuals in our training set. And in this context, the only way to use some of other methods was through low dimension features.

4.2 Feature Engineering

In this task, we have extracted features in different ways or for different data sets in order to perform different models.

- *Feature extraction by considering only popular pages.*

In order to obtain popular pages, we looked at pages liked by more than 500 users. Doing so we had more than 150 pages remaining. Each user in our relation data set were represented by a one hot encoding of those pages. Further, we retained only 49 pages according to information theory based feature selection. The graph in the figure 8 shows how the retained pages were discriminating. As we can see in the figure 8, they have pretty much the same discriminatory behavior, and still remain redundant even if they have been chosen by information gain with penalized redundancy criterion. In addition, ignoring less popular pages implied ignoring thousands of users who liked only such pages. After obtaining low performances on this features, we moved to the next extraction technique.

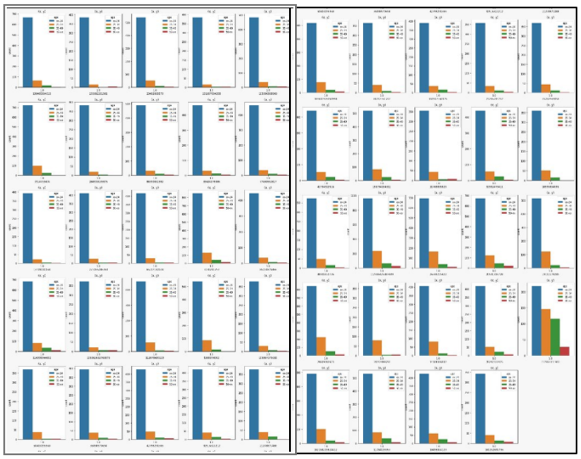

Fig. 8. Age group distribution within each popular page (Note that for this graph the blue, orange, green and red colors mirrors respectively the cohorts xx-24, 25-34, 35-49 and 50-xx)

- *Grouping pages by popularity scale.*

In this technique, we extracted 6 features, corresponding to categorizing pages in the popularity scale ,namely : pages liked by a single user, pages liked by 2 till 10 users, pages liked by 11 till 50 users, pages liked by 51 till 100 users, pages liked by 101 till 500 users, pages liked by more than 500 users. The discriminatory behavior of those 6 features are summarized in figure 9 box-plots (Note that we have considered users who have liked at most 50 pages within each of the 6 categories). In our observation we observe the same discriminatory nature i.e the young generation liking more pages.

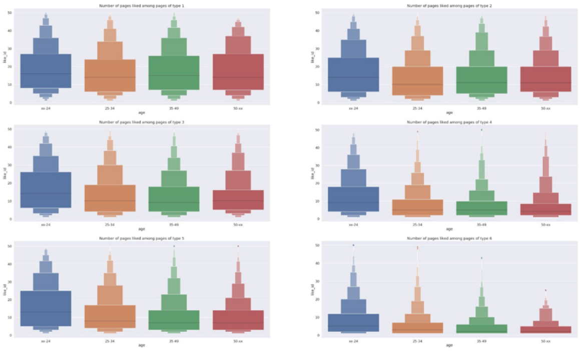

Fig. 9. box plots summarizing the number of pages liked by a user within each of the 6 types of pages *(from the left to the right, from top to bottom, pages liked by a single user, pages liked by 2 till 10 users, pages liked by 11 till 50 users, pages liked by 51 till 100 users, pages liked by 101 till 500 users, pages liked by more than 500 users)*

- *Looking at text features in NRC and LIWC data sets.*

Next, we moved to NRC and LIWC features. We trained the classifiers, using either one or two data sets or both. But again, all the methods performed here (KNN, MLP, Naive Bayes, SVM, Decision Tree) led to very low accuracy rate, even when we mixture of text features with relation set.

4.3 Results

After experimenting we finalized the relation data as our data set for this classification task we realized. In order to handle the huge number of feature which comes along with this dataset we chose MLP model as our final model. The choice of this model relied on the fact that it is more a convenient method when we have too many features comparing to sample size. In our case, we had more than 500000 one hot encodings of page likes for just 8500 individuals in our training set. And in this context, the only way to use some of other methods was through low dimension features.

It is clear why we chose the MLP model given relation features, there's another question on why did we choose to use the relation features. There is a intuition-based and a feature-engineering based answer. The intuition was that the features available on image data sets are not relevant for this task. And we can think that people in the same generation are more likely to target similar pages. As for the feature-engineering based answer, after trying to use NRC and LIWC features, we always obtained low performances comparing to relation data set hence we used relation features as our final data set. The results on the data set - that is, one hot encodings of all page likes - are given in table 2 in the conclusion section. We obtained an accuracy rate of 65.6% with the model (MLP), which is 6% higher than the baseline performance.

5 PERSONALITY PREDICTION TASK

5.1 Methodology

This task consists of predicting the big five personalities. We decided to build 5 unrelated regression models. For the personality part, we used a Multilayer Perceptron regression model with two hidden layers of 100 neurons each, and four linear regression models.

5.2 Feature engineering

The most basic way to find out about one's personality is in the way that person express himself/herself. That is the assumption/intuition which made us decide to use the messages of the users as our dataset for this task. As mentioned before, the messages were already preprocessed into LIWC and NRC features. We applied PCA on the 90 features in order to reduce the dimension. It turned out that the first component contains 99% of the explained variance. This is telling us that those 90 features were highly correlated. The figure 10 is the plots of the second principal component against the first principal component, were we have highlighted the range of values of two different personality traits (openness and conscientiousness). Again, this support the fact that the original features were highly correlated, and that applying PCA could help our method.

5.3 Results on the regression models

We tried two different models: MLP on the original features and linear regression on the new PCA features.
For the MLP model, we tuned the number of hidden layers and the number of neurons per layer. It gave us results that were already better than the baseline. For the linear regression, in order to select the best number of principal component to use, we used the number of components that minimize the RMSE on the validation set. The figure 11 shows the validation curves for four of the five traits. It turns out that one of the personality trait did not give a better result on the linear regression, so we kept the MLP model.

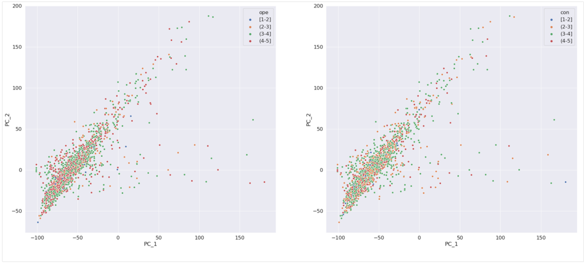

Fig. 10. Plots of the two best principal components

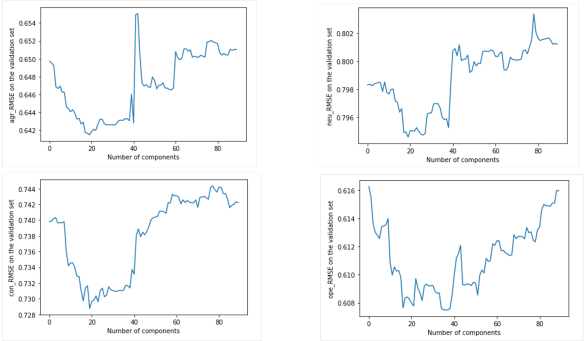

Fig. 11. Validation RMSE against the number of principal components, for four of the five traits

6 CONCLUSION

From the results we obtained in the below tables, we can affirm that the data provided are relevant in order to predict the age, the gender and the personality.
We observed that some data sets are more helpful towards particular classification task, for example the oxford features suits for gender classification task and the relation data suits more for age classification task.

For gender classification task:

- It would be interesting to see how setting the facial hair feature for female gender creates the algorithmic bias. Getting a quantized metric for such bias will help to make the system more fair and equal.
- One of the interesting thing which can be done in the future work is to get the actual profile pictures (not just the facial points) and predict the age of the user. The features like wrinkle sink on the forehead could be very helpful in the age prediction task.

With more time, we would have liked to try improving our models, especially for the personality task, since our results were sometimes

very close to the baseline.
We had two approaches in mind:

- Our first approach would be to try new methods, like a convolutional network regression, which we discovered only near the end of the session. In fact, it seems to handle well the type of data that are highly correlated, in which relations between the features need to be made.
- Our second approach would be to add new features: the likes. The assumption is that people with similar personality might have similar habits. For example, introverted people and extroverted people.

The tables below sum up our results.

Table 1. Accuracy for the gender classification task

	MLP + NB	MLP	NB	SVM	KNN	Baseline
Gender	**87%**[1]	84.7%	88%[2]	78.8%[2]	78.9%[2]	59.1%[1]

Table 2. Accuracy for age classification task

	MLP	NB	SVM	Baseline
Age	**65.6%**[1]	37%	30%	59.4%[1]

Table 3. RMSE of the regressors

	MLP Regressor	**Linear Regressor**	Baseline
Ope	0.608	0.607 (0.651)[1]	0.652[1]
Con	0.728	0.728 (0.722)[1]	0.734[1]
Ext	0.797 (0.783)[1]	0.802	0.788[1]
Agr	0.645	0.641 (0.657)[1]	0.665[1]
Neu	0.795	0.794 (0.794)[1]	0.796[1]

[1] These results were obtained on the test set (see scoreboard on the course website)
[2] These results were based on users having profile pictures