User Profiling in Social Media

Alaka Sukumar
University of Washington
alakas@uw.edu

Jayant Keswani
University of Washington
jk0805@uw.edu

Sai Prajna Vadlamani
University of Washington
vadlas@uw.edu

Vikhyati Singh
University of Washington
singhv3@uw.edu

ABSTRACT

The phenomenal rise of social media in recent years presents new opportunities and challenges to mine actionable patterns from a large amount of data to understand user behavior and to meet users information needs. Social media services like Facebook have emerged as important platforms for large-scale information sharing and communication in fields such as marketing, journalism, and public relations. Profiling this information for use in recommender systems, advertisements, e-commerce applications is attracting the attention of many researchers in many disciplines with the aim of understanding the behavior of social media users using machine learning techniques.The goal of this project is to build a prediction system which predicts the age, gender and personality traits such as Openness, Conscientiousness, Extroversion, Agreeableness, Emotional Stability. The features need to be recognized given the user's Profile image, Status Updates and Page Likes information of Facebook users.

Keywords

Classification; Regression; Naive Bayes; RMSE; Personality; features; labels.

1. INTRODUCTION

Social networks have become widely-used, and the information collected from this social media provide a valuable insight into individual behavior, experiences, opinions and interests.Personalized systems used in domains such as, e-learning, information filtering, collaboration and e-commerce could greatly benefit from profiling user data.
This paper provides the design and implementation details of an automatic recognition system for predicting the age, gender and personality of Facebook users. This system returns as output the age, personality traits and gender of the user, when provided with input, the status updates and profile picture of that user as input. Our research has two interconnected objectives: (1) to predict age and gender of Facebook users given the status updates as input (2) the relevant personality-related indicators that are explicitly or implicitly present in Facebook user data.

The Big five personality traits is established as the most popular research model which identifies the Big Five dimensions: Openness to experience, Neuroticism, Extraversion, Agreeableness and Conscientiousness. We hypothesized that increasing the relevance of what is included in the model, and considering features drawn from a variety of sources may lead to better performance of the classifiers under investigation. Our research is currently focused on building models for prediction of personality traits and prediction of age and gender of Facebook users.

2. RELATED WORK

There have been various research efforts in mining social media information. Among them, the work by Dejan Markovikj, Sonja Gievska,Michal Kosinski and David Stillwell deals with the study of personality reflected in user's Facebook activities[1]. The study explores the feasibility of modeling user personality based on a proposed set of features extracted from the Facebook data.The study builds a classification model using Support Vector Machines (SVM)[2] and their more efficient and optimized versions, Simple Minimal Optimization (SMO)[3] and Boost algorithms (MultiBoostAB and AdaBoostM1)[4], and at each stage refining the classification mechanisms by considering better sampling of features based on the Pearson correlation coefficient.An employment of ranking algorithms attributed significant performance gains for the SMO classifier to predict personality traits.

"Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures", a research by Fabio Celli, Elia Bruni and Bruno Lepri[5], make use of exploited a bag-of-visual-words technique to extract features from pictures. The goal of this research was to automatically recognize personality traits and interaction styles from Facebook profile images using the Scale Invariant Feature Transform (SIFT), one of the most popular and effective feature extraction technique used for object recognition[6]. The research by Dr. Huan Liu, from Arizona State University , on "Advancing the Frontier of Social Media Mining" [7] focuses on verifying if information provided across various social media sites belong to the same individual.

Personality-related features with linguistic cues is another approach to build models based on social network activities(Mairesse Oberlander and Nowson) [8]. They conduct a set of experiments to examine whether automatically trained models can be used to recognise the personality of unseen subjects.The correlation between users' social network activity and personality has been the focus of several studies in

the last decade (Bai, Zhu, and Cheng 2012; Golbeck, Robles, and Turner 2011; Bachrach et al. 2012). Personality traits of the Chinese most popular social network RenRen users were analyzed in (Bai, Zhu, and Cheng 2012)[9]. Decision trees have shown the best results, yielding 69-72 percent accuracy, for a combination of features related to users network activity along with affective linguistic features extracted from statuses and blog posts.

3. INITIAL EXPLORATORY WORK

As a part of our initial work, we read and understood the research by Golnoosh Farnadi and team in their work "Recognising Personality Traits Using Facebook Status Updates"[10]. The goal was to predict personality traits of Facebook users using Facebook status updates. The idea was based on the fact that since more than one trait can be present in the same user, for each trait train a binary classifier that separates the users displaying the trait from those who do not. Four kinds of numeric features were used in the experiments - Linguistic Inquiry and Word Count (LIWC) features, Social Network features, Time Related features and Other features such as total number of statuses per user. The three learning algorithms used on these features were Support Vector Machine with a linear kernel (SVM), Nearest Neighbor with k=1 (kNN)[11] and Naive Bayes (NB) [12].

We chose Python as the programming language after careful evaluation of the pros and cons of python, java and R for data analysis. We familiarized ourselves with Python's machine learning libraries such as numpy, pandas, scikit-learn etc.

Next, we built a baseline classifier that generated random values for Age, Gender and Personality Traits. We then modified the baseline classifier to output the average values instead of random values.

4. DATA MODEL

We are presented with a training dataset of 9500 labelled instances. The training dataset includes the floowing information:

1. **Profile Information** A csv file with userid, age, gender and personality scores of all 9500 users
2. **Text** Text files with status updates of all the users. There are 9500 such text files.

Figure 1: Gender distribution for the given public data

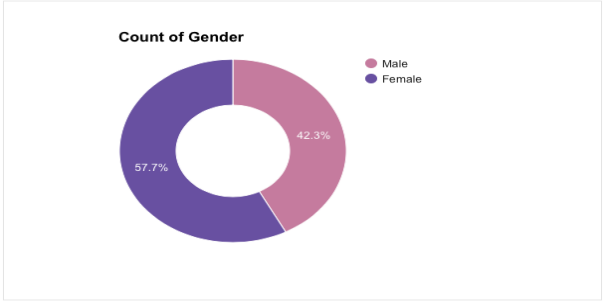
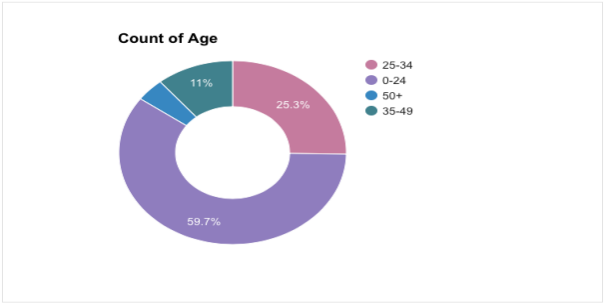


Figure 2: Age distribution for the given public data



3. **Image** Profile pictures of the users, there are 9500 such images
4. **Relation** A csv file with rows of the format "user-id", "like-id", indicating that the user with "user-id" liked the page with "like-id"
5. **LIWC** The Linguistic Inquiry and Word Count (LIWC) consist of features related to standard counts , psychological processes , relativity, personal concerns , and linguistic dimensions,totalling 81 features which are mapped to each user.
6. **NRC** A csv file with rows of NRC emotion features

A test dataset with data of 334 Facebook users without labels is also provided.The test dataset also contains the profile csv file, the text files, the profile pictures, LIWC , NRC and relations.

5. COMPUTING ENVIRONMENT

Ubuntu Workstation 14.04.3 LTS,
2GB hard disk and 2GB RAM.
Python Version 2.7
scikit-learn version 0.17.1
R version 3.2.2

6. METHODS EXPLORED

There are numerous methods that we tried and tested before we finalized on the final classifier. We explain in brief about the various techniques we tried during the course of this research.

1. **Textblob Library** We predicted age and gender by using the Naive Bayes Classifier from the textblob library. A bag of words was created from the status updates of users and the labels provided in the profile.csv. We passed this bag of words to the Naive Bayes classifier. The main challenge with this technique was that the training was extremely slow and we were not able to train the entire 9500 train files which resulted in low accuracy.
2. **Logistic and Linear Regression using Images** In this technique, each image loaded using SKIMAGE. We used the Google Tensor Flow library to predict the images as this allows to specify model as graph of nodes. We used Logistic Regression for Gender Prediction and Linear Regression for Age Prediction.For each test image we generated : 200 * 200 * 1 grey

scale value. Regression Techniques computes Weight Matrix and Bias. We then use the computed Weight Matrix and Bias to predict the score

$$Y = Wx + B$$

(X is input vector) We obtained an accuracy of 60% for Age Prediction and 58% for Gender Prediction using the public test data.

3. **Personality Prediction using NRC and LIWC** In this technique a data frame has been created by loading NRC features of training data and another data frame by loading profile.csv of the training data.Both these data frames have been merged to form a complete data frame for training data.Using the .fit() and .predict() methods, the model is trained and used for predicting personality traits respectively. Obtained RMSE values higher than baseline during prediction and hence we did not include this technique in the final model.

4. **Personality Prediction using Random Forests** Merged LIWC features and profile.csv of training data. Used the Random Forest Regressor to predict the personality traits and compute RMSE. The RMSE obtained were satisfactory and higher than baseline values.

5. **Age, Gender and Personality Prediction using Page Likes** Using the Facebook Graph API giving likeID as the input we got all the information about the pages such as name, category, ID, description, weekly active users and many more. Out of total 79 attributes we chose the categories as the feature to be used. There exists a one to many mapping. Each user is mapped to 240 categories (sparse matrix). And then the random forest technique is used for prediction of age, gender and personality prediction.

6. **Naive Bayes without using python libraries** We also implemented Naive Bayes with using libraries.The data was pre-processed by dividing text files based on age and gender and training data was modelled by calculating word frequency and probabilities(Naive method)

7. FINAL METHODOLOGY

We use supervised learning methods to build our model as the training dataset is labelled. The final classifier uses Classification for predicting age and gender using status updates as input and and Regression for predicting the personality traits using LIWC features as input. We implemented the Naive Bayes algorithm for age and gender prediction and Linear Regression for personality prediction.

7.1 Classification

We implemented Naive Bayes classifier to predict age and gender using Facebook status updates as input. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features.It is one of the most basic text classification techniques with various applications in email spam detection, personal email sorting, document categorization,language detection and sentiment detection. In a text classification problem, we will use the words or tokens of the document in order to classify it on the appropriate class.This means that in order to find in which class we should classify a new document, we must estimate the product of the probability of each word of the document given a particular class (likelihood), multiplied by the probability of the particular class (prior). After calculating the above for all the classes of set C, we select the one with the highest probability.

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \; p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

.

7.1.1 Implementation Details

We used Naive Bayes classifier from (Natural Language Toolkit)nltk packages, to predict age and gender using Facebook status updates. In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong (naive) independence assumptions between the features. Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s, and remains a popular (baseline) method for text categorization[10]. Naive Bayes is a conditional probability model: Given a data instance (X) to be classified, represented by a vector of n features ($x_1, ..., x_n$), it assigns class instance probabilities: $P(C_k|X) = P(C_k|x_1, ..., x_k)$. Applying the bayesian theorem, this class probability can be derived by the following equation:

$$P(C_k|X) = \frac{P(C_k)P(X|C_k)}{P(X)}$$

The above equation of assigning class probabilities can also be interpreted as following:

$$Posterior = \frac{Prior * likelihood}{evidence}$$

While implementing various algorithms for age and gender prediction using text, we observed, for a very simple algorithm, naive bayes gives much better results than many other algorithms.NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. Our data preprocessing included conversion into lower case, removal of redundant white spaces, stemming, removal of stop words, correction of spaces between commas etc. We noticed substantial improvement in accuracy by using porter stemmer. This is especially true for small training datasets. Stemming reinforces signal for certain features. NLTK library has variety of stemmer and stop words list for each language. Using stemmer and stop word removal helped improve our training accuracy.

7.1.2 Experiments and Results

On the hidden test data, our classifier obtained an accuracy of 0.62 for Age and 0.74 for Gender. Out of the all the implementations that we conducted experiments with, the

native Naive Bayes classification model proved to provide better prediction accuracy as shown in table 1.

Figure 3: Comparison of various implemented techniques for text classification

Method	Algorithm	Accuracy for Age	Accuracy for Gender
Using Textblob library	Naïve Bayes	0.59	0.41
Using Images as input	Logistic Regression	0.6	0.59
Using nltk library	Naïve Bayes	0.62	0.74

7.2 Regression

We implemented Linear Regression to predict the personality traits.Linear regression is the most basic and commonly used predictive analysis. Regression estimates are used to describe data and to explain the relationship between one dependent variable and one or more independent variables. Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest.

7.2.1 Implementation Details

We predicted personality traits using Linear Regression. The input features are LIWC features. A data frame has been created by loading LIWC features of training data and another data frame by loading profile.csv of the training data has also been created. Both these data frames have been merged to form a complete data frame for training data. Along with the LIWC features, age and gender have also been considered in the feature selection. Since users from test data do not have age and gender, they are initialized to default values i.e, 'xx-24' for age and 'female' for gender. This data is fed to the linear regression model after a 10 fold validation. Here, out of the 9500 files, 70% is used for training and 30% for testing.

Using the .fit() and .predict() methods, the model is trained and used for predicting personality traits respectively. Each of these personality traits, which are in the form of a numpy array, are outputted to a csv file for each user. Along with the predictions, Root Mean Square Error(RMSE) for each personality is also calculated and printed as output.

7.2.2 Experiments and Results

Figure 4 summarizes the results (RMSE) obtained for the Big 5 personality traits using LIWC features with a 70-30% split on training data.

7.3 Summary of our Final Model

So our final classifier predicts age and gender using the Natural language tool kit (NLTK) library's Naive Bayes Classifier for text classification and provides a good accuracy of 0.62 for age and 0.74 for gender. The input was the status updates of users. The approach was to count each document word frequency and all document word frequency using nltk, perform pre-processing functions such as standard tokenization, stemming and stop words removal using nltk and using Naive Bayes classifier to predict the age and gender.

Figure 4: Personality Prediction using LIWC features with 70-30% split on training data

Gender	Age	ope(RMSE	ext(RMSE)	con (RMSE	agr(RMSE)	neu(RMSE)
No	No	0.64	0.8	0.8	0.82	0.81
Yes	Yes	0.69	0.86	0.73	0.7	0.71
No	Yes	0.93	0.97	1.01	0.87	0.81
Yes	No	0.85	0.93	0.88	0.78	0.79

Our final classifier uses Linear Regression to predict the Big 5 personality traits using LIWC features and provides RMSE error values of 0.66,0.80,0.80,0.70,0.77. We merged LIWC and profile.csv to create a data frame for the classifier. Age and Gender have also been considered in the feature selection and the RMSE was computed.

The Split Information is as follows: Training Data = 70%, Test Data = 30% of training data.

8. CHALLENGES

1. During personality prediction using LIWC and NRC, we encountered null values in profile.csv of test data that lead to errors,we replaced the null values with zeroes to rectify the error.
2. During text classification using status updates, we faced challenges due to memory constraints while training the model using text blob library.
3. Handling unicode related errors was challenging.
4. Image processing was very memory and compute intensive because of large feature space.
5. Training was very slow during image classification.
6. Getting the information about the Page like using Facebook Graph API was very time consuming

9. CONCLUSIONS

In this paper, we have summarized the results of our extensive experiments on mining data from Facebook users and have presented some good techniques for building and testing the models. We built regression and classification models and extensively evaluated our models for correctness and improving accuracy. For personality prediction, linear regression with LIWC showed better performance over random forest. Using age and gender as attributes, reduced the RMSE during personality prediction. Naive Bayes classifier is the best method to do text classification.

10. FUTURE WORK

Future research into this area could be focused on attempting to improve the mechanism of the comparative probability age classifier, to see if the performance may improve as it did for the gender classifier. Other future work could look into utilization of frequency distribution information

in ensemble classifiers to achieve improved prediction accuracy overall.Augmenting the personality models with a more qualitative features like pages, groups, events can be experimented.

11. ACKNOWLEDGMENTS

Our thanks to Professor Martine De Cock for giving us the opportunity to conduct this project on Machine Learning and many thanks to Golnoosh Farnadi for her suggestions throughout the project.

12. REFERENCES

[1]Dejan Markovikj,Sonja Gievska,Michal Kosinski,David Stillwell,*Mining Facebook Data for Predictive Personality Modeling*,Proceedings of AAAI International Conference on Weblogs and Social Media (ICWSM), 2013.

[2]https://en.wikipedia.org/wiki/Support_vector_machine

[3]https://en.wikipedia.org/wiki/Sequential_minimal

[4]http://www.boost.org/doc/libs/?view=category_Algorithms

[5]Fabio Celli,Elia Bruni,Bruno Lepri, *Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures*, Proceedings of ACM international conference on multimedia , pp 1101-1104
[6]https://en.wikipedia.org/wiki/Scale-invariant_feature_transform

[7] Prof. Huan Liu,*Advancing the Frontier of Social Media Mining*,BYAC 270, 2014

[8]Francoise Mairesse,Marilyn A. Walker ,Matthias R. Mehl,Roger K. Moore *Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text*,Journal of Artificial Intelligence Research 1

[9]Shuotian Bai, Tingshao Zhu, Li Cheng,*Big-Five Personality Prediction Based on User Behaviors at Social, Network Sites*,April 2012

[10]https://en.wikipedia.org/wiki/Naive_Bayes_classifier