Author Profiling in English Tweets

Filip Zelić, Borna Sirovica, Ivan-Dominik Ljubičić

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{filip.zelic, borna.sirovica, ivan-dominik.ljubicic}@fer.hr

Abstract

In this paper, we present our approach to the author profiling task, a student project for Text Analysis and Retrieval course. Given a set of tweets by the same person, the task aims at identifying age, gender and personality traits of that person. We address age and gender prediction as a classification task and a personality prediction as a regression problem. We experimented with Support Vector Machine for classification and regression and other machine learning algorithms using a variety of custom designed features as well as features extracted from publicly available resources.

1. Introduction

Author profiling distinguishes between classes of authors by studying their sociolect aspect, i.e., how language is shared or how an author can be characterized from a psychological viewpoint. This information helps in identifying profiling aspects such as gender, age, native language, or personality type.

Author profiling is a problem of growing importance, among others for applications in forensics, security, and marketing. However, social profiling still remains a less-explored topic, even though the exponential growth of social networks increased its importance even further. While there were several publications that were trying to predict some demographical information such as gender, age, and native language (Argamon and Shimoni, 2003), (Peersman and Vaerenbergh, 2011), as well as the personality type, and to perform author profiling in general (Argamon and Schler, 2009), the real push forward was enabled by specialized competitions such as the PAN shared tasks on author profiling (Pardo and Inches, 2013), (Rangel and Daelemans, 2014), (Rangel and Daelemans, 2015), which ran in 2013–2015.

This paper presents our approach to the author profiling task. The task focused on predicting an author's demographics (age and gender) and the big five personality traits (McCrae and Costa, 2008) (agreeable, conscientious, extroverted, open, stable) from a set of tweets by the same target author.

This paper is organized as follows: Section 2 gives a detailed description of dataset used for training and testing of our model. Section 3 presents general approach and methodology used in the experiments, including a description of the preprocessing, the features, and the learning algorithms we used. Section 4 presents and discusses our results and provides some deeper analysis. Finally, Section 5 concludes and points to possible directions for future work.

2. Dataset

The dataset we used consisted of English tweets from 152 users in XML format along with one *truth.txt* file with age, gender and the Big Five personality traits labels for each user (PAN, 2015). For labeling age, the following classes were considered: 1) 18—24, 2) 25—34, 3) 35—49, 4) 50+ and gender was labeled as male (M) or female (F). The distribution of the gender and age labels in the corpus is reported in Table 1. For the case of gender classes the corpus was balanced with 50% of the tweets labeled as male and other half as female, but regarding age the distribution was skewed due to the lower number (around 22%) of the older users (labels *35—49* and *50+*) and higher number (around 78%) of the

younger users (labels *18—24* and *25—34*).

Table 1: Distribution of Twitter users with respect to the age and gender labels in the corpus.

Trait	Label	Number of users
Age	18—24	58
	25—34	60
	35—49	22
	50+	12
Gender	Male	76
	Female	76

Regarding personality traits normalized numeric rating in [-0.5,0.5] range was given for each of the following properties: extroverted, stable, agreeable, conscientious, and open. The mean for each trait is reported in Table 2.

Table 2: Mean values of the Big Five personality traits in the corpus.

Personality trait	Mean value
Extroverted	0.16
Stable	0.14
Agreeable	0.12
Conscientious	0.17
Open	0.24

3. Approach

Along this section, we describe the steps we took to prepare features for the training of our supervised machine learning models. In subsection 3.1 we explain the preprocessing step in which we cleaned the data in order to get better performances for stylometric features. Subsection 3.2 explains extraction of various features organized by their category and the last subsection 3.3 describes classification and regression models we used tackling the task of author profiling.

3.1. Preprocessing

Preprocessing is done by creating a document for each user joining all his/hers tweets from the dataset. After the creation, the document is initially stripped of XML tags and cleared of all twitter specific characteristics such as hashtags, @replies as well as URLs from the text. While doing this step we save the count of mentions, hashtags and URLs for later use in feature modeling. When the preprocessing step is done, features need to be extracted.

3.2. Feature extraction

As mentioned in the introduction, we used many different self-designed features and features extracted from publicly available resources for our machine learning algorithms.

Lexicon Features. We used several lexicons, both manually crafted and automatically generated:

- **NRC Word-Emotion Association Lexicon** (Mohammad, 2013): one dictionary for each of the eight primary human emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust as well as dictionaries for positive and negative emotion word classification.
- **Internet slang word lexicon** : personally made dictionary with most common words younger people like to use on the Internet.
- **Frequent male and female words lexicon** (Schwartz and Ungar, 2013): dictionaries containing words most likely to be used by male or female person. Based on word usage frequency per million words.
- **Frequent male and female function words lexicon** (Kiprov, 2015) : dictionaries containing function words, such as 'this','the','she', or 'not', most likely to be used by male or female person.

For each lexicon, we found the number of occurrences of lexicon terms in all tweets for each user, normalized by the total number of words in tweets. We instantiated separate features for the different lexicons.

Twitter-specific Features. We used the following Twitter-specific features:

- **Hashtags:** the number of hashtags;
- **URLs:** the number of URLs posted;
- **User mentions:** the number of mentions of users using the pattern @username;
- **Emoticons:** the number of emoticons used in tweets.

All of the above counts are normalized by the total number of available tweets for the target user; so they could be viewed as "average number per tweet".

Orthographic Features. We used the following orthographic features:

- **Letter case:** the number of upper-case words and upper-case characters;
- **Character flooding:** the number of redundant character reduplication;
- **Word length:** average word length;
- **Tweet length:** average tweet length;
- **Specific characters:** usage of specific characters, such as the number of occurrences of the exclamation points, question marks or apostrophes.

Term-level Features. We used the following term-level features:

- **n-grams:** TF-IDF matrices of unigrams and trigrams;
- **F-score:** measure of contextuality and formality based on the frequency of the part of speech (POS) usage in a text (f.x below means the frequency of the part-of-speech x):

$$F = 0.5 * [(f.noun + f.adj + f.prep + f.art) - (f.pron + f.verb + f.adv + f.int) + 100]$$

- **POS tags:** frequency of certain POS tags such as nouns, verbs, prepositions, determiners, interjections, adverbs, adjectives etc.

3.3. Supervised learning models

Supervised machine learning models are used with hyperparameter optimization using grid search and k-fold cross validation on 70% of training data. Regarding classification task for age and gender, we tested the following classification models:

- Support Vector Machine Classification (SVC) with linear kernel and RBF kernel
- Logistic Regression
- Random Forest

On the other hand regarding regression task for personality traits, following regression models were tested:

- Support Vector Machine Regression (SVR) with linear kernel and RBF kernel
- Linear Regression

Values of features were scaled using StandardScaler from sklearn python library in order to avoid complications that can arise in the classification stage when features with numeric values that differ a lot.

4. Evaluation

After extraction of features and normalization of feature values, we trained models from Section 3.3 on 70% of data and evaluated them on other 30% of data trough multiple iterations in which train and test data were shuffled. In this process we tried combinations of many features from Section 3.2 for each subtask. Best feature combinations are stated in the next few paragraphs.

Regarding age subtask best scores were achieved using combinations of twitter-specific and orthographic features along with TF-IDF scores of frequent trigrams. SVM model with RBF kernel proved to be the best choice model wise.

In Gender classification subtask best results were obtained with the usage of male and female based lexicons together with orthographic features and F-score measure. SVM model with RBF kernel was proven to be the best choice for model selection also. The results of these two classification subtasks are given in the Table 3.

Table 3: Age and Gender prediction accuracy

Subtask	Classifier	Accuracy score
Gender	SVM (RBF kernel)	77.2 %
Age	SVM (RBF kernel)	76.1%

Personality traits subtask was modeled as a regression problem. For every trait we chose the most appropriate feature combination which in general included word-emotion based lexicons as well as the F-score measure, POS tags frequencies and orthographic features such as upper-case word count, average tweet length and exclamation overload count. To measure accuracy of personality traits prediction root mean squared error (RMSE) measure was used. Lowest errors between predicted and actual scores were obtained using SVM Regression model. Results of this task are given in following Table 4.

Table 4: Personality traits prediction accuracy

Personality trait	Regression model	RMSE
Extroverted	SVR (RBF kernel)	0.142
Stable	SVR (RBF kernel)	0.170
Agreeable	SVR (RBF kernel)	0.145
Conscientious	SVR (RBF kernel)	0.153
Open	SVR (RBF kernel)	0.156

5. Conclusion and future work

This paper proposes a supervised machine learning technique for solving the author profiling problem of determining author's age, gender and Big Five personality traits trough feature extraction from tweet corpus.

Tackling this challenging task we combined a large number of various features in order to train and evaluate our machine learning models. The main problem was that we could not evaluate our models on the most real case scenario because we did not find a publicly available evaluation dataset so the models were evaluated on 30% of the training dataset.

Even though evaluation showed good results for age gender and personality trait prediction (Rangel and Daelemans, 2015), we feel that there is a lot more to gain. For example, in the future work we could develop a more sophisticated approach for personality trait identification, considering more specific features and preprocessing for each personality trait separately. We could also experiment with descriptive LSA (Latent Semantic Analysis) and discriminative SOA (Second Order Attributes) features analyzed in paper (Álvarez Carmona and Escalante, 2015).

References

Pennebaker Argamon, Koppel and Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM*, 2:119–123.

Argamon and Shimoni. 2003. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17:401—-412.

Nakov-Koychev Kiprov, Hardalov. 2015. Experiments in author profiling—notebook for pan. *CLEF PAN 2015*.

McCrae and Costa. 2008. A contemplated revision of the neo five-factor inventory. *The SAGE handbook of personality theory and assessment*.

Mohammad. 2013. Research in computational linguistic. *National Research Council Canada*.

PAN. 2015. Pan15-author-profiling-training-dataset-english. *https://github.com/dmincu/Author-Profiling/tree/master/pan15-author-profiling-training-dataset-2015-03-02/pan15-author-profiling-training-dataset-english-2015-03-02*.

Koppel-Stamatatos Pardo, Rosso and Inches. 2013. Overview of the author profiling task at pan 2013. *Working Notes for CLEF 2013 Conference*.

Daelemans Peersman and Vaerenbergh. 2011. Predicting age and gender in online social networks. *Workshop on Search and Mining User-generated Contents, SMUC '11*, pages 37–44.

Chugur-Potthast Trenkmann-Stein-Verhoeven Rangel, Rosso and Daelemans. 2014. Overview of the 2nd author profiling task at pan 2014. *CLEF 2014 Evaluation Labs and Workshop*.

Potthast-Stein Rangel, Rosso and Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. *CLEF 2015 Labs and Workshops*.

Kern-Dziurzynski Ramones-Agrawal-Shah Stillwell-Seligman Schwartz, Eichstaedt and Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *In PLOS ONE*, 8.

Montes-y-Gómez Villaseñor-Pineda Álvarez Carmona, López-Monroy and Escalante. 2015. Author profiling task notebook for pan at clef 2015. *INAOE's participation at PAN'15*.