Author profiling

Lovre Mrčela, Marko Ratković, Ante Žužul

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{lovre.mrcela, marko.ratkovic, ante.zuzul}@fer.hr

Abstract

The goal of this project was to profile an author by analyzing a set of texts written by them, and then determining degree of each the Big Five personality traits. In addition, gender and age–group for each author are derived as well. The dataset was collected from twitter profiles, in English, Italian, Spanish and Dutch languages. Approach was based on *tf-idf*, considering occurences of trigrams. The implementation is done in Pythong programming language, using *nltk* and *sklearn* libraries.

1. Introduction

Author profiling deals with problem of describing someone's personality, by means of extracting information from their writing style. Personality can be described using five traits (the so-called *"Big Five personality traits"*), which are: extraversion, stability, agreeableness, conscientiousness and openness to experience. Degrees of each trait range from -0.5 (indicating the total opposite), to 0.5 (indicating the exact match).

Provided with degrees of the five traits, it is possible to determine author's gender and age–group, via classification based on a model trained on previously labelled data. In this project, we used the linear SVC and Gaussian naive Bayes models for the classification into gender and age–group, and the linear regression with squared error measure for determining the degrees of personality traits. The training set we used was a collection of twitter posts in English, Spanish, Italian and Dutch authors, ranging from around 35 authors in Dutch to 150 in English, each author's file containing about 100 posts.

2. Approach

In this section, the methods of our approach are thoroughly explained. First, the preprocessing of input text is carried out, and weighted vector of trigrams (three consecutive letters) is obtained. Then, from preprocessed text some additional feature vectors, which were reasonably expected to be discriminative, are extracted. Finally, gender and age–group classification and personality traits regression models are trained on extracted features, and final results are compared for various parameters.

2.1. Text preprocessing

For the rest of the process to be optimal, some sort of text preprocessing needs to be done on the raw input data. The input data we use is given in *xml* format, so the first step in preprocessing was to parse the actual sentences from the *xml* structure. When that is done, following steps are also applied:

- *urls* to other sites are substituted with an URL tag, and usernames (when referenced in replies) are substituted with a REPLY tag,
- all the text is converted to lower case because we don't deal with capitalization of words, only with words themselves;
- more than 3 repetitions of the same letter are reduced to 3 letters, so that the words like *"coooool"* (5 repetitions) and *"coooooool"* (7 repetitions) are both treated as the same word, but distinctly from *"cool"*, because while we want to take repetitions into account, we would like to ignore the quantity of repeated letters (see the Section 2.2.);
- stop words (for that particular language) are deleted from the text, because they are considered unsignificant for author profiling.

Each three consecutive letters are grouped into trigrams, and weighted vector of trigrams is obtained, using *tf–idf* weighting scheme. The extracted trigram weighted vector is used as one feature. More features are then extracted from preprocessed text, as described in the next subsection.

2.2. Additional feature extraction

In addition to weighted vector of trigrams, we decided to investigate some further characteristics of the written corpora, which were expected to be discriminative for the gender and/or age–group. Here is the list of cosidered additional features, and explanation for each of them:

- **number of emoticons:** the average number of emoticons used in a post (e.g. :) , <3; not considering each emoticon distinctly but all of them in total),
- **number of consecutive long repetitions of letters:** as mentioned before, we count only occurences of repetitions longer than 3 letters, not the length of repetitions themselves – these repetitions most of the time do not have constant number of letters, even for the same author, or the same post, so it is a better approach to take into account only instances of repetitions;
- **number of replies:** the average number of replies to another user per each post,
- **number of hashtags:** the average number of hashtags per post,

- **number of exclamation marks:** the average number of exclamation marks (!) per post – each exclamation marks is counted, as we considered that, opposed to the consecutive repetition of letters, repeated exclamation marks do indicate author's stronger emotion to a some degree.
- **average length and standard deviation of posts:** we were inspecting average post length, as we presume it may also be correlated with age–groups;
- **average length and standard deviation of words:** as above, but considering just words.

It was expected for some of the features to be present in a greater degree in some subpopulations compared to the other (i.e. younger vs. older, male vs. female). The obtained results respective to each feature are shown in the section 4..

2.3. Gender and age–group classification

For the gender and age–group classification subproblem, following approaches were considered:

- Logistic Regression
- Naive Bayes Classifier
- Decision Tree Classifier
- Random Forest Classifier
- SVC (using *rbf*, linear, poly- and sigmoid kernels)

The best result was obtained by using SVC with linear kernel, compared to other methods.

2.4. Personality traits regression

For the personality traits regression, following approaches were considered:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- SVR (using various kernels)

After testing each method, the best results turn out to be obtained by using SVR and linear regression.

3. Testing

Due to the lack of access to the official testing dataset (because of an ongoing competition), the official training dataset was divided into a subset for training (around 70%) and a subset for testing (around 30%).

The model hyperparameters for gender and age–group classifiers are optimized using *k–fold* cross–validation.

4. Results

For the age–group and gender classification, baseline error is represented as score of Naive Bayes Classifier. For the personality traits, we defined baseline error as root mean–square error of average of all features. Precision, recall, F1 micro score, and macro score measures for each language are shown in table **??**.

5. Conclusion

Experimenting with various models and features, we obtained results similar to other published works (in our case, tested on reduced training set). Unfortunately, we were not able to test our solution on the official data due to the data not yet having been released.

The possible upgrade of this work would be researching application of Latent Semantic Indexing, as it may further improve detection of author personal traits.

References

G. B. Johnson and W. W. Howells. 1974. Title title title title title title title title title title. *Journal journal journal*.

G. B. Johnson, W. W. Howells, and A. N. Other. 1976. Title title title title title title title title title title. *Journal journal journal*.

Table 1: Overview of results per language.

Language	Precision	Recall	F1	Macro score
English *Italian* *Spanish* *Dutch*	0.71787	0.64035	0.65214	0.78261