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 occurrences of trigrams. The implementation is done in Python 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 labeled 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. Of these four sets, English and Spanish are labeled with age–group, while the Italian and Dutch sets are not.

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 character 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 characters (see the Section 2.2.);
- stop words(**?**) for that particular language are deleted from the text, because they are considered insignificant 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 considered 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 characters:** as mentioned before, we count only occurrences of repetitions longer than 3 characters, not the length of repetitions themselves – these repetitions most of the time do not have constant number of characters, 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 with respect to each feature are shown in the section 4..

The final feature set was obtained by selecting *n* best features, where *n* is also hyperparameter that needs to be optimized as well as the model. We select *n* best features using the ANOVA F–value[1].

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 results for age–groups were obtained using SVC with linear kernel, and for binary classification of gender, the Gaussian Naive Bayes.

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.

All of these models are implemented in *sklearn* library, which we are using in our project solution.

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 (70%) and a subset for testing (30%).

From above mentioned models, optimal model and hyperparameters were selected by using *10–fold* cross–validation. Criterion for classifier (which was also used in PAN contest(**?**)) was accuracy score. Aside from accuracy, we also used precision, recall, and F1 measures (micro and macro, for multi–class classification) (**?**). For the regressor, we used the root–mean–square error.

In the absence of official testing dataset, it was obligatory to set baseline score as a referent measure. The baseline was set by using dummy models. Baseline classifier always gives the most frequent class, and baseline regressor always outputs the mean value. Thus, achieved results can be put into a more real perspective.

4. Results

Further observation of additional features values shows there is correlation between some features and age–groups and/or gender of the author. For example, average post length tends to be longer for older users than for younger users, for both the English and Spanish corpora. Number of user replies in average is also greater for older users than for younger users, for both languages. Average number of hashtags per posts seems to slightly increase towards older users in English corpus, however it is not in linear correspondence with age in Spanish corpus.

Also, there is correlation between the average number of emoticons per posts considering the gender of user: in the English, Italian and Dutch corpora, female users in average tend to use up to three times as many emoticons than male users. However, in Spanish corpus the situation is reversed: male users in average tend to use in more emoticons than female users. Same goes for number of exclamation marks: female users in all languages use on average more exclamation marks than male users.

Some features seem not to have any correspondence to either age–group or gender, for example the post length deviation. We can see those correlations in the tables 1 (age–groups) and 2 (gender).

Precision, recall, F1 micro score, and macro score measures for each language are shown in tables 3 (age–groups), 4 (gender), and 5 (personality traits).

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 approach of Latent Semantic Analysis, as it may further improve detection of author personal traits.

[1] http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

Table 1: Overview of additional features values for each age–group, per language.

Language	English				Spanish			
Age–group	18-24	25-34	35-49	50-XX	18-24	25-34	35-49	50-XX
Post length	60.714	85.853	86.680	93.753	75.728	85.246	92.804	101.991
Post length deviation	29.656	29.401	29.532	32.192	31.377	31.234	30.719	29.200
Word length	4.908	5.984	6.312	6.013	4.935	5.295	5.647	5.409
Word deviation	3.493	4.726	5.088	4.452	3.356	3.882	4.230	3.963
Emoticon count	0.057	0.064	0.046	0.038	0.135	0.104	0.053	0.030
Hashtags	0.127	0.658	0.267	0.514	0.168	0.340	0.259	0.231
Character repetitions	0.040	0.012	0.017	0.003	0.065	0.022	0.030	0.022
Exclamation marks	0.137	0.207	0.195	0.527	0.183	0.244	0.257	0.276
User replies	0.492	0.540	0.632	1.293	0.579	0.715	0.818	0.854

Table 2: Overview of additional features values for gender, per language.

Language	English		Spanish		Italian		Dutch	
Gender	Female	Male	Female	Male	Female	Male	Female	Male
Post length	76.786	77.222	86.030	86.949	91.555	87.513	77.442	77.239
Post length deviation	29.711	29.764	30.767	31.131	32.908	30.594	29.574	30.829
Word length	5.529	5.718	5.328	5.281	5.898	6.153	5.255	5.229
Word length deviation	4.187	4.385	3.916	3.786	4.202	4.705	3.489	3.476
Emoticons count	0.075	0.039	0.082	0.102	0.214	0.072	0.119	0.072
Hashtags	0.380	0.394	0.357	0.190	0.540	0.700	0.424	0.120
Character repetitions	0.026	0.020	0.041	0.025	0.008	0.006	0.026	0.016
Exclamation marks	0.275	0.133	0.252	0.221	0.228	0.168	0.271	0.121
User replies	0.641	0.548	0.745	0.698	0.725	0.556	0.647	0.909

Table 3: Overview of results of age–group classification per language. Baseline scores are given for comparison

Language	Accuracy	Precision	Recall	$\mathrm{F1}^{(\mathrm{micro})}$	$\mathrm{F1}^{(\mathrm{macro})}$
English	0.782	0.717	0.640	0.652	0.782
English (baseline)	0.326	0.081	0.250	0.122	0.326
Spanish	0.933	0.975	0.900	0.924	0.933
Spanish (baseline)	0.600	0.150	0.250	0.187	0.600

Table 4: Overview of results of gender classification per language. Baseline scores are given for comparison

Language	Accuracy	Precision	Recall	F1
English	0.956	0.888	1.000	0.941
English (baseline)	0.347	0.347	1.000	0.516
Spanish	1.000	1.000	1.000	1.000
Spanish (baseline)	0.400	0.400	1.000	0.571
Italian	1.000	1.000	1.000	1.000
Italian (baseline)	0.416	0.000	0.000	0.000
Dutch	1.000	1.000	1.000	1.000
Dutch (baseline)	0.454	0.454	1.000	0.625

Table 5: Overview of RMSE of personality traits regression per language. Baseline scores are given for comparison

Language	Extraversion	Stability	Agreeableness	Conscientiousness	Openness
English	0.122	0.179	0.153	0.140	0.130
English (baseline)	0.164	0.235	0.182	0.167	0.155
Spanish	0.080	0.143	0.103	0.155	0.131
Spanish (baseline)	0.123	0.220	0.149	0.211	0.183
Italian	0.048	0.123	0.067	0.086	0.099
Italian (baseline)	0.136	0.166	0.116	0.162	0.162
Dutch	0.087	0.112	0.129	0.064	0.041
Dutch (baseline)	0.137	0.197	0.155	0.115	0.116

References