Automatic Profiling of Twitter Users Based on Their Tweets

Notebook for PAN at CLEF 2015

Octavia-Maria Şulea[1,2] and Daniel Dichiu[1]

[1] Bitdefender Romania
[2] Center for Computational Linguistics, University of Bucharest
mary.octavia@gmail.com, ddichiu@bitdefender.com

Abstract In this paper we go through our approach at solving the PAN Author Profiling task. We introduce a novel way of computing the type/token ratio of an author and show that, although strong correlations have been observed between high extroversion and low type/token ratios in the past, this ratio is not necessarily a strong indicator of extroversion. Since the text of a person is influenced by all 7 features (gender, age, and big five personality traits) that are required to be automatically identified in this task, we used this ratio, along with Term frequency-Inverse document frequency (*tf-idf*) matrices, in all 7 subtasks and all 4 corpora and obtained good results.

1 Introduction

While the importance of age and gender is a more familiar notion in user or author profiling, automatic personality detection is a relatively new task [10]. Since many correlations between personality traits and consumer preferences have been reported ([5], [9]), a natural interest arose in the automatic detection of personality on social media networks in the last few years, especially on the micro-blogging site, Twitter, where the privacy setting for its users' posts and activity is by default *public* ([7], [3]). Since the main activity of Twitter users involves language (tweets), and since many correlations have been identified between lingusitic features of a text and personality traits of its author [4], the idea of automatically detecting the personality of Twitter users based on their tweets is only natural. In what follows, we will describe our approach to PAN's third Author Profiling task [8], discuss our cross-validation results and briefly compare them with the results obtained after the final testing.

2 Our Approach

For all datasets and subtasks, the estimators, the parameter search function, the cross-validation strategy, and some of the feature extractors we used were from the scikit-learn module for python [6]. For the processing of the other features, we also used the nltk module [1]. This implementation choice of python modules was motivated by the swiftness with which prototyping can occur. The two classification tasks (for gender and age) were carried out using LinearSVC() [2] while the 5 regression tasks (for the personality traits), using Ridge(). In order to have balanced classes during cross-validation, we used StratifiedKFold() with the number of folds set to 5. The best parameters for the estimators were found using RandomizedSearch().

For features, we tried several approaches, but eventually settled on using two: first, the *tf-idf* matrix at character level, with various n-gram ranges and parameter tuning, depending on the language and subtask, and second, the type/token ratio of a user or *verbosity* rate. These two features were combined using scikit-learn's FeatureUnion().

The *tf-idf* scores were extracted using scikit-learn's TfidfVectorizer(). This vectorizer was applied either on all tweets of one user put together, or on each tweet pertaning to one user. More precisely, in the sparse matrix created by the TfidfVectorizer(), the columns represented, in both cases, all the character n-grams extracted from all the tweets in one of the four datasets, while each line represented either all tweets of one user concatenated, or one tweet of a user. Our cross-validation results, which will be presented further, showed that the former method was consistently more appropriate for the classfication tasks and the latter, for regression. An intuitive answer would be that gender and age specific features change less often, while personality traits may influence each tweet.

The verbosity ratio of a user was only inspired by the type/token ratio and is not one per se, since distinguishing between a linguistic manifestation of a conceptual type (*bicycle* in sentence 1.a), and its token (*bicycle* in sentence 1.b), implies deep semantic analysis which is far from trivial with today's tools in Natural Language Processing.

(1) Type/token distinction

a. The *bicycle* is more popular now. *Type*

b. The *bicycle* is in the garage. *Token*

What we did to echo the idea of a type/token ratio was to compute, for each user, the ratio between the total number of *unique* stems and the total number of words used after applying stemming. From this ratio we excluded stop words. Stemming was done using the nltk implementation of the Snowball algorithm since it offered a version for each of the four languages present in this year's task. For stopwords lists, we used nltk.corpus.stopwords. The motivation for using this feature was the often observed correlation between extroversion and type/token ratio [4].

However, our preliminary analysis, by computing both Pearson and Spearman correlation coefficients on verbosity ratios versus personality scores, showed no clear-cut linear relationships. The fact that Spearman correlation coefficient was better than the Pearson correlation coefficient only goes to show that the relationship is rather a monotonic one than linear. Below are the top three statistically significant correlation scores on all corpora and all personality scores, computed with python's scipy.stats package. The plots were drawn using seaborn python module.

For the Dutch corpus (figure 1 on page 3), we found that there was a -0.46 Pearson correlation with a p-value < 0.001 and -0.49 Spearman rank correlation with a p-value < 0.001 between verbosity ratios and openness scores. Also, given a verbosity ratio, males tended to have higher openness scores than females.

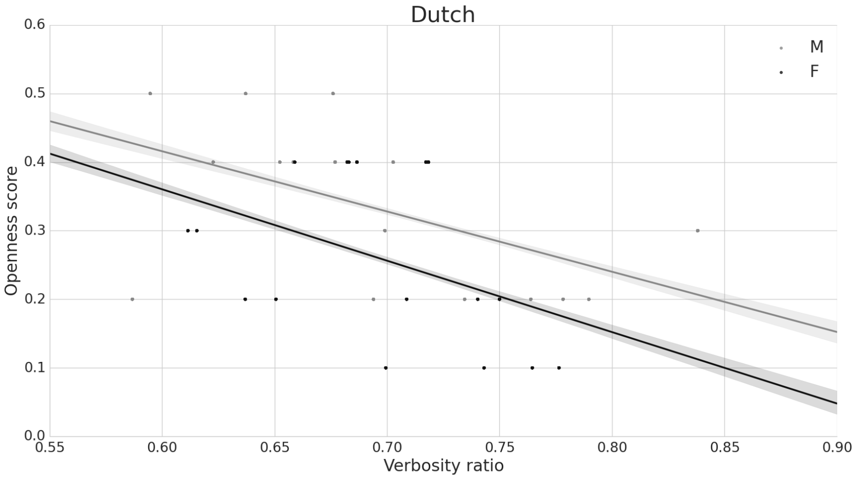

Figure 1: A somewhat negative correlation between a Dutch person's openness score and his/her verbosity ratio.

Also for the Dutch corpus (figure 2 on page 4), we found that there was a -0.34 Pearson correlation with a p-value < 0.001 and -0.45 Spearman rank correlation with a p-value < 0.001 between verbosity ratios and neuroticism (stable) scores. Regarding gender separation, males tended to be more stable at a given verbosity ratio.

For the Italian corpus (figure 3 on page 4), we found that there was a -0.33 Pearson correlation with a p-value < 0.001 and -0.40 Spearman rank correlation with a p-value < 0.001 between verbosity ratios and agreeableness scores. Apparently, on average, Italian females were more agreeable than Italian males at a given verbosity ratio.

We also present the results of verbosity ratios for the classification tasks.

In the English training corpus (table 1 on page 3), across all age groups, males had a slightly higher verbosity ratio than females. We also observed that verbosity ratios increased slightly with age, across both genders.

Table 1: Verbosity ratios on English

Gender	Median	Mean	Std
female (all ages)	0.6731	0.6584	0.0887
male (all ages)	0.683	0.68	0.083
18-24 (both genders)	0.6763	0.6650	0.0782
25-34 (both genders)	0.6864	0.6704	0.0921
35-49 (both genders)	0.6756	0.6737	0.0903
50-xx (both genders)	0.6927	0.6769	0.0989

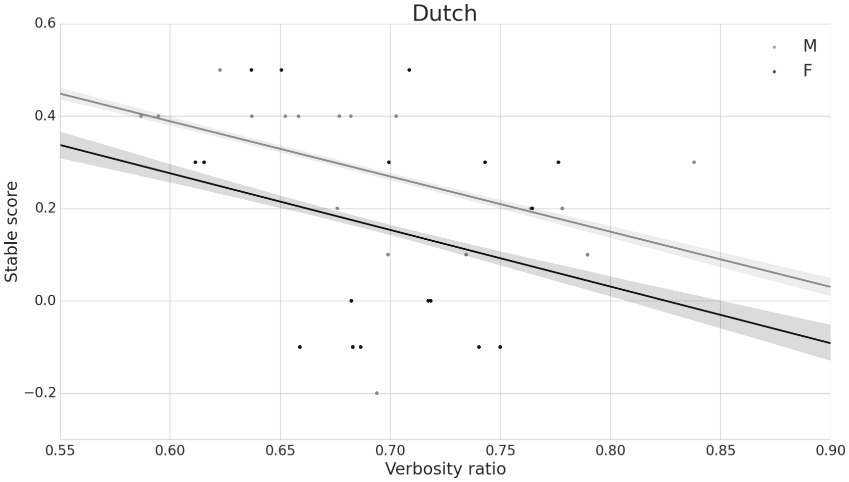

Figure 2: A somewhat negative correlation between a Dutch person's neuroticism score and his/her verbosity ratio.

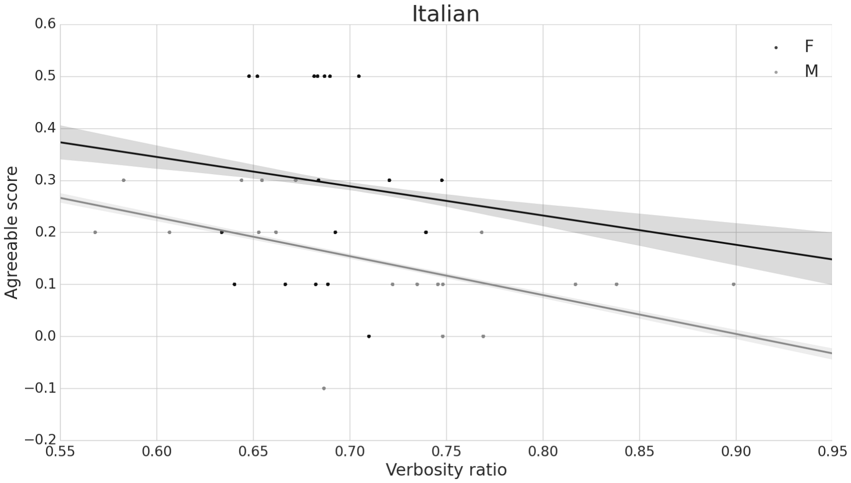

Figure 3: A somewhat negative correlation between an Italian person's aggreableness score and her/his verbosity ratio.

The difference betweem male and female verbosity ratios was minimal on the Spanish training corpus (table 2 on page 5). However, we observed a larger difference when it came to age groups, with the highest verbosity ratio being for age group 25-34 (with a median of 0.70) and the lowest for age group 35-49 (with a median of 0.67).

Table 2: Verbosity ratios on Spanish

Gender	Median	Mean	Std
female (all ages)	0.6937	0.6817	0.0483
male (all ages)	0.6901	0.6855	0.0623
18-24 (both genders)	0.6844	0.6858	0.0645
25-34 (both genders)	0.7016	0.6873	0.0514
35-49 (both genders)	0.6693	0.6787	0.0528
50-xx (both genders)	0.6859	0.6715	0.0614

A similar difference we also observed on the Italian corpus (table 3 on page 5). Female users tended to have a lower verbosity ratio (with a median of 0.69), while males had a median verbosity ratio of 0.72.

Table 3: Verbosity ratios on Italian

Gender	Median	Mean	Std
female	0.6870	0.6891	0.0328
male	0.7223	0.7117	0.0874

As for the Dutch training corpus (table 4 on page 5), the difference between male and female verbosity ratios was again minimal, with a difference between medians of under 2 percentage points.

Table 4: Verbosity ratios on Dutch

Gender	Median	Mean	Std
female	0.6995	0.6967	0.0502
male	0.6822	0.6934	0.0694

Given these inconclusive findings, we decided to use a combination of *tf-idf* on character n-grams with verbosity scores, which improved cross-validation results over models based on the same features taken separately.

3 Cross-Validation Results

Table 5: TfidfVectorizer parameters and results on English

Subtask	Range	Max-df	Min-df	Sublinear_tf	Vocab.	CV result	Result
gender	$\overline{1,3}$	0.75	0.17	TRUE	3211	78.94%	76.76%
age	$\overline{3,5}$	0.98	0.14	FALSE	13677	75.65%	78.87%
stable	$\overline{2,6}$	N/A	N/A	TRUE	773075	0.1825	0.1951
agreeable	$\overline{2,6}$	N/A	N/A	FALSE	773075	0.1411	0.1396
extroverted	$\overline{2,6}$	N/A	N/A	TRUE	773075	0.1359	0.1318
conscientious	$\overline{2,6}$	N/A	N/A	TRUE	773075	0.131	0.1297
open	$\overline{2,6}$	N/A	N/A	TRUE	773075	0.1193	0.1246

Table 6: TfidfVectorizer parameters and results on Spanish

Subtask	Range	Max-df	Min-df	Sublinear_tf	Vocab.	CV result	Result
gender	$\overline{2,6}$	0.85	0.15	FALSE	20649	88%	87.5%
age	$\overline{1,3}$	0.82	0.07	FALSE	6540	73%	75%
stable	$\overline{2,6}$	N/A	N/A	TRUE	563605	0.1812	0.1816
agreeable	$\overline{2,6}$	N/A	N/A	FALSE	563605	0.1478	0.1501
extroverted	$\overline{2,6}$	N/A	N/A	TRUE	563605	0.1517	0.1703
conscientious	$\overline{1,3}$	0.94	0.07	TRUE	431	0.1137	0.1559
open	$\overline{2,6}$	N/A	N/A	FALSE	563605	0.1421	0.1417

Comparing the cross-validation results to the final test results, we can see signs of overfitting only in some of the cases in which we used relatively more features. Overall, our models generalized well when the number of features was smaller. LinearSVC() and Ridge() allowed us to use sparse matrices, which meant we did not have to transform to dense matrices (which would have occupied too much memory) or reduce dimensions (which is a computationally expensive operation).

As we stated before, we concatenated each user's tweets for the classification tasks, while for the regression tasks we used each individual tweet. This led, on average, to a smaller vocabulary for the classification tasks.

On the English corpus (table 5 on page 6), our system over-fitted slightly on the gender, stable and open tasks. On the Spanish corpus (table 6 on page 6), our system over-fitted slightly on extroverted and conscientious tasks. On the Dutch corpus (table 7 on page 7), our system over-fitted slightly on extroverted and conscientious tasks.

The biggest difference between cross-validation and test results was on the Italian corpus (table 8 on page 7), where our system over-fitted on all tasks, but extroverted. The biggest overfit was for the gender task, with a difference of 15 percentage points between cross-validation results and test corpus results.

Table 7: TfidfVectorizer parameters and results on Dutch

Subtask	Range	Max-df	Min-df	Sublinear_tf	Vocab.	CV result	Result
gender	$\overline{1,3}$	0.95	0.17	FALSE	3663	76.47%	84.38%
stable	$\overline{1,3}$	N/A	N/A	TRUE	17015	0.1592	0.1405
agreeable	$\overline{2,6}$	N/A	N/A	FALSE	237795	0.123	0.1114
extroverted	$\overline{2,6}$	N/A	N/A	TRUE	237795	0.1075	0.131
conscientious	$\overline{2,6}$	N/A	N/A	FALSE	237795	0.0964	0.1147
open	$\overline{3,5}$	N/A	N/A	TRUE	127649	0.0915	0.0846

Table 8: TfidfVectorizer parameters and results on Italian

Subtask	Range	Max-df	Min-df	Sublinear_tf	Vocab.	CV result	Result
gender	$\overline{5,7}$	0.72	0.1	TRUE	31745	78.94%	63.89%
stable	$\overline{3,5}$	N/A	N/A	TRUE	165628	0.1502	0.1913
agreeable	$\overline{3,5}$	N/A	N/A	FALSE	165628	0.1227	0.122
extroverted	$\overline{2,4}$	N/A	N/A	FALSE	74719	0.1191	0.1141
conscientious	$\overline{3,5}$	N/A	N/A	FALSE	165628	0.101	0.114
open	$\overline{2,6}$	N/A	N/A	FALSE	307144	0.1298	0.1438

4 Conclusions

Based on our results, we conclude that a combination of simple features like *tf-idf* and verbosity ratios obtain reasonable results that generalize well. Comparing our approach across all corpora, we found that this solution worked best as a regressor for the Dutch corpus and as a classifier for the Spanish corpus. We found that the best *tf-idf* features are those at character-level ngrams, with ngram ranges of up to $\overline{2,6}$. Above this threshold, the system seemed to overfit. We also found that there is at best a monotone relationship between verbosity ratios and personality scores. Nevertheless, combining them with other, many-dimensional features, like *tf-idf* matrices, improves results and generalizes well.

References

1. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc. (2009)
2. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (June 2008)
3. Golbeck, J., Robles, C., Edmondson, M., Turner, K.: Predicting personality from twitter. In: SocialCom/PASSAT. pp. 149–156. IEEE (2011), `http://dblp.uni-trier.de/db/conf/socialcom/socialcom2011.html#GolbeckRET11`
4. Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research (JAIR pp. 457–500 (2007)
5. McCrae, R.R., Costa, P.T.: Personality in Adulthood: A Five-Factor Theory Perspective (2nd ed.). New York: Guildford (2003)

6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (Oct 2011)
7. Quercia, D., Kosinski, M., Stillwell, D., Crowcroft, J.: Our twitter profiles, our selves: Predicting personality with twitter. In: Proceedings of the Third International Conference on Social Computing (SocialCom) and the Third International Conference on Privacy, Security, Risk and Trust (PASSAT). pp. 180–185. IEEE (Oct 2011), http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6113111&tag=1
8. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) CLEF 2015 Labs and Workshops, Notebook Papers. CEUR-WS.org (2015)
9. Roozmand, O., Ghasem-Aghaee, N., Nematbakhsh, M., Baraani, A., Hofstede, G.: Computational modeling of uncertainty avoidance in consumer behavior. International Journal of Research and Reviews in Computer Science pp. 18–26 (April 2011)
10. Vinciarelli, A., Mohammadi, G.: A survey of personality computing. T. Affective Computing 5(3), 273–291 (2014), http://dx.doi.org/10.1109/TAFFC.2014.2330816