An Author Profiling Approach Based on Language-dependent Content and Stylometric Features

Notebook for PAN at CLEF 2015

Alberto Bartoli, Andrea De Lorenzo, Alessandra Laderchi,
Eric Medvet, and Fabiano Tarlao

DIA - University of Trieste, Italy
bartoli.alberto@univ.trieste.it, andrea.delorenzo@units.it, alessandra.laderchi@gmail.com,
emedvet@units.it, fabiano.tarlao@phd.units.it

Abstract We describe the approach that we submitted to the 2015 PAN competition [5] for the author profiling task[1]. The task consists in predicting some attributes of an author analyzing a set of his/her Twitter tweets.
We consider several sets of stylometric and content features, and different decision algorithms: we use a different combination of features and decision algorithm for each language-attribute pair, hence treating it as an individual problem.

1 Problem statement

A problem instance consists of a tuple $\langle D, L \rangle$, where D is a set of tweets written by the same author and L is a value of enumerated type that describes the language of the tweets—English, Spanish, Italian, or Dutch.

The author profiling consists in generating, given a problem instance, the value for several attributes with respect to the author of the tweets: gender, age group (only for English and Spanish), and 5 personality traits. Age group is an enumerated value among the following: 18–24, 25–34, 35–49 or $\geqslant 50$. The 5 *personality traits* are widely accepted characteristics used to describe human personality (also known as Big Five [7]): extroversion, neuroticism, agreeableness, conscientiousness, and openness to experience. For each trait, the attribute value consists of a score in $[-0.5, +0.5]$.

A set of solved problem instances (the *training set*) is available in which, for each problem instance $\langle D, L \rangle$, the tuple of the attributes values is provided.

The effectiveness of a method for author profiling is assessed using a *testing set* of solved problem instances. In particular, the effectiveness is assessed separately for each attribute as follows: the attribute values generated by the method for the problem

[1] During the competition we discovered several opportunities for fraudulently boosting the accuracy of our method during the evaluation phase. We will describe these opportunities in a future report. We notified the organizers which promptly acknowledged the high relevance of our concerns and took measures to mitigate the corresponding vulnerabilities. The organizers acknowledged our contribution publicly. We submitted for evaluation an honestly developed method—the one described in this document—that did not exploit such unethical procedures in any way.

instances in the testing set are compared against the actual values and the comparison outcome is expressed in terms of accuracy for gender and age, and in terms of Root-mean-square error (RMSE) for the personality traits.

2 Our approach

We chose to handle the prediction of each attribute for each language as an individual problem: in particular, we consider gender and age group prediction as 2 classification tasks and personal traits prediction as 5 regression tasks. Since we had tweets written in four languages and we had to predict age groups for those written in English and Spanish only, we hence considered 26 different problems.

We propose a machine learning approach based on a number of different *stylometric* and *content* features which are processed by one among three different decision algorithms—we used SVM and random forests as classifiers and regressors. We carried out an extensive experimental campaign for systematically assessing a large number of the possible combinations, through *leave-one-out* cross validation on the available training data.

2.1 Training set analysis and repetitions

During preliminary analysis, we noticed that the training set included some subsets of problem instances for which L and the solution were the same, i.e., the attributes values for all the problem instances in a subset were the very same, despite being D different. We call *repetitions* those problem instances. We argued that the tweets of the problem instances in each of those subsets were authored by the same person. For this reason, we decided to build a new training set by replacing each of those subsets with a single problem instance in which D is the union of all the tweet sets of the subset—i.e., we merged the repetitions. Table 1 shows the sizes of the training set portions corresponding to each language before and after merging repetitions. We later experimentally verified that this transformation did affect the learned classifiers and regressors.

Language	Original	Merged
English	152	83
Spanish	100	50
Italian	38	19
Dutch	34	18

Table 1. Number of problem instances in the original training set and in the new training built by merging repetitions.

2.2 Features

The feature extraction procedure requires a language-dependent *dictionary* in which words are grouped according to their prevalent topic (e.g., “money”, “sports”, or “religion”) or their function (e.g., “prepositions”, “articles”, or “negations”). To this end,

we used an English dictionary similar to the one used by LIWC [4]. For the other 3 languages, we proceeded as follows. For Spanish and Dutch, we built the dictionary by automatically translating the English dictionary with Google Translate. For Italian, we manually built the dictionary, by using the English one as guideline. Moreover, for each language, we augmented the dictionary with a new category of words ("chat acronyms") containing the top fifty most popular chat acronyms exposed on NetLingo[2].

The feature extraction procedure is also based on the notion of *automatic tweet*, that we define as follows. We determined a set of ordered sequences of $n = 1, \ldots, 4$ words, that we call *templates*, based on an analysis of the full training set:

1. we automatically extracted from the full training set all tweets starting with the same ordered sequence of n words;
2. we automatically constructed a set including all word sequences that were the starting sequence of at least 3 different tweets;
3. we manually analyzed each sequence and retained only those which appeared to be the beginning of an automatically-generated tweet.

We say that a tweet is an automatic tweet if its first words correspond to a template. Table 2 provides some examples of templates, along with the presence or absence of corresponding automatic tweets of different languages in the training sets.

Template	EN	ES	IT	NL
`# Move más reciente`		✓		
`Photo:`	✓	✓	✓	
`I'm at`	✓	✓	✓	
`I liked a`	✓		✓	
`I favorited a`	✓	✓	✓	
`Ik vind een`	✓			✓
`#in`	✓		✓	
Total number of templates	29	8	12	1

Table 2. Some examples of templates and the languages for which at least one automatic tweet with that template were found. The first row corresponds to a template found only in Spanish problem instances, while the other rows are templates found in problem instances of multiple languages. The last row contains, for each language, the count of templates for which at least one automatic tweet with that template was found.

The feature extraction procedure is as follows. Given a problem instance $\langle D, L \rangle$, we denote by D_M the set of tweets obtained by D by removing all the automatic tweets. We extract several numerical features from each problem instance: the value of all (except of 3) features is obtained by averaging the corresponding computation outcomes on the tweets in D or D_M—the remaining three feature values are computed on the whole D and/or D_M. For ease of presentation, we group conceptually similar features together; the full list is given in Table 3.

[2] http://www.netlingo.com/top50/popular-text-terms.php

Stylometric These features tend to capture the structural properties of a tweet in a way largely independent of both the language and the specific semantic content; therefore, they are not based on the dictionaries. Stylometric features are computed on tweets in D_M: the reason is because we assume that automatic tweets are not really representative of the tweet writing style of the author.

Content These features are based on the dictionaries categories related to word topic and are computed on tweets in D: the reason is because we assume that the content of automatic tweets is indeed informative of the author profile.

Hybrid These features are based on the dictionaries categories related to word function and are computed on tweets in D_M.

2.3 Feature selection

Past studies on author profiling report several correlations between gender, age, personality traits and writing style. In particular, [6] showed that stylometric features are more predicitve than content features for determining the gender, and viceversa for the age group, but the combination of both stylometric and content features can offer better results. In [3], the authors provided a list of correlations between some LIWC and non-LIWC features and the five personality traits. We constructed 40 different feature groups based on this knowledge and we assessed each of the resulting feature groups as described in the next section.

2.4 Classifier and regressor

We decided to build a different model for each language-problem pair, for a total of 26, as described in Section 1. We explored the usage of SVM [2] and Random Forest [1] with different configurations, as these methods can be used both as classifiers and as regressors. In particular, we considered:

- *svm*: SVM with default gaussian kernel and $C = 1$;
- *rf500*: Random Forest with 500 trees;
- *rf2000*: Random Forest with 2000 trees.

3 Analysis

As described in the previous sections, we considered 40 sets of features and 3 classifiers/regressors. We systematically assessed the effectiveness of all the 120 resulting combinations by means of a *leave-one-out* procedure applied on the training set, separately for each language-attribute pair. That is, for each language-attribute pair, set of features, and classifier/regressor, (i) we built the subset T of the problem instances of the training set with that language, (ii) we removed one element t_0 from T, (iii) we computed the values for the features set on the problem instances in T and trained the classifier/regressor, (iv) we applied the trained classifier/regressor to the problem instance t_0 and compared the generated answer against the known one. We repeated all but first steps $|T|$ times, i.e., by removing each time a different element, and computed

	Feature name	Description
stylometric	allpunc	Number of `.,:;`
	commas	Number of `,`
	exclmar	Number of `!`
	questma	Number of `?`
	parenth	Number of parenthesis
	numbers	Number of numbers
	wocount	Number of words
	longwor	Number of words longer than 6 letters
	upcawor	Number of uppercase words
	carrret	Number of carriage returns (`\n`, `\r`, `\r\n`)
	atmenti	Number of `@` mentions
	extlink	Number of links
	hashtag	Number of #
	posemot	Number of positive emoticons
	negemot	Number of negative emoticons
	emotico	Number of emoticons
	emotiyn	Presence of emoticons in D (binary feature)
content	moneywo	Number of words in the "money" category
	jobword	Number of words in the "job or work" category
	sportwo	Number of words in the "sports" category
	televwo	Number of words in the "tv or movie" category
	sleepwo	Number of words in the "sleeping" category
	eatinwo	Number of words in the "eating" category
	sexuawo	Number of words in the "sexuality" category
	familwo	Number of words in the "family" category
	frienwo	Number of words in the "friends" category
	posemwo	Number of words in the "positive emotion" category
	negemwo	Number of words in the "negative emotion" category
	emotiwo	Number of words in the "positive emotion" or "negative emotion" category
	swearwo	Number of words in the "swear words" category
	affecwo	Number of words in the "affective process" category
	feeliwo	Number of words in the "feeling" category
	religwo	Number of words in the "religion" category
	schoowo	Number of words in the "school" category
	occupwo	Number of words in the "occupation" category
	autotwe	Automatic tweets ratio, i.e., $\frac{\lvert D \setminus D_M \rvert}{\lvert D \rvert}$
	autweyn	Presence of automatic tweets in D (binary feature)
hybrid	fsipron	Number of words in the "I" category
	fplpron	Number of words in the "we" category
	ssipron	Number of words in the "you" category
	selfref	Number of words in the "self" category
	negpart	Number of words in the "negations" category
	asspart	Number of words in the "assents" category
	article	Number of words in the "articles" category
	preposi	Number of words in the "prepositions" category
	pronoun	Number of words in the "pronoun" category
	slangwo	Number of words in the "chat acronyms" category

Table 3. Features list.

the performance of the method in terms of the indexes defined in Section 1. Finally, we chose, for each language-attribute pair, the best performing combination, in terms of accuracy or RMSE, as appropriate for that attribute. The resulting configurations are summarized in Table 4.

In order to provide a synthetic baseline, we built 3 baseline methods using each of the 3 classifiers/regressors with all the features. The results, obtained by means of the same leave-one-out procedure, are shown in Table 5.

It can be seen from Table 4 that our procedure lead us to chose a different configuration of classifier/regressor and features set for each language-attribute pair. There could be several reason to explain that. First, every language has its own writing rules and culture, so it is possible that a middle aged English man could not have the same interests and the same writing style of a middle aged Italian man. Second, the Spanish, Dutch, and Italian dictionaries we used were not as good as the LIWC English one. Finally, the number of problem instances in the training set was not the same for every language, and so was the number of tweets in the instances within each language subset.

References

1. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
2. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 27 (2011)
3. Golbeck, J., Robles, C., Edmondson, M., Turner, K.: Predicting personality from twitter. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. pp. 149–156. IEEE (2011)
4. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count (liwc): A computerized text analysis program. Mahwah (NJ) 7 (2001)
5. 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 Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2015), http://www.clef-initiative.eu/publication/working-notes
6. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs 6, 199–205 (2006)
7. Soldz, S., Vaillant, G.E.: The big five personality traits and the life course: A 45-year longitudinal study. Journal of Research in Personality 33(2), 208–232 (1999)

L	Attribute	Class./Regr.	Chosen features set
EN	Gen	rf2000	commas negemot exclmar
	Age	rf2000	allpunc commas exclmar questma parenth numbers wocount longwor upcawor carrret atmenti extlink hashtag posemot negemot emotico autotwe
	Ext	svm	wocount questma parenth familwo
	Neu	svm	selfref fsipron chatacr affecwo emotiwo hashtag posemot pronoun wocount
	Con	rf500	extlink longwor numbers hashtag fsipron selfref
	Agr	svm	questma atmenti allpunc ssipron article longwor jobword chatacr
	Ope	rf2000	commas extlink hashtag exclmar questmar parenth wocount ssipron negpart article feeliwo moneywo jobword eatinwo familwo negemwo religwo
ES	Gen	svm	allpunc commas exclmar questma parenth numbers wocount longwor upcawor carrret atmenti extlink hashtag posemot negemot fsipron fplpron ssipron selfref negpart asspart article preposi pronoun slangwo moneywo jobword sportwo televwo sleepwo eatinwo sexuawo familwo frienwo posemwo negemwo affecwo feeliwo
	Age	svm	extlink hashtag numbers sleepwo sexuawo
	Ext	rf2000	longwor carrret questma preposi autweyn emotico
	Neu	rf2000	posemot ssipron exclmar selfref extlink
	Con	rf500	extlink longwor numbers hashtag fsipron selfref affecwo emotiwo
	Agr	svm	allpunc commas exclmar questma parenth numbers wocount longwor upcawor carrret atmenti extlink hashtag posemot negemot + fsipron fplpron ssipron selfref negpart asspart article preposi pronoun slangwo moneywo jobword sportwo televwo sleepwo eatinwo sexuawo familwo frienwo posemwo negemwo swearwo religwo
	Ope	rf2000	autotwe hashtag preposi wocount religwo
IT	Gen	rf500	asspart fsipron selfref exclmar extlink hashtag emotiyn
	Ext	svm	allpunc wocount hashtag questma
	Neu	rf2000	commas longwor fplpron chatacr autweyn
	Con	svm	commas extlink hashtag exclmar questmar parenth wocount ssipron negpart article feeliwo moneywo jobword eatinwo familwo negemwo religwo
	Agr	svm	posemot exclmar moneywo hashtag pronoun autweyn
	Ope	svm	negpart hashtag atmenti exclmar longwor
NL	Gen	rf2000	negemot upcawor preposi
	Ext	svm	questma atmenti allpunc ssipron article longwor jobword chatacr extlink autweyn
	Neu	rf2000	atmenti preposi longwor emotiyn
	Con	svm	hashtag questma exclmar atmenti posemot wocount extlink longwor
	Agr	svm	atmenti commas exclmar hashtag autweyn emotiyn
	Ope	svm	negpart hashtag atmenti exclmar longwor

Table 4. Chosen classifier/regressor and features set for each language-attribute pair.

		Baselines			
L	Attribute	svm	rf500	rf2000	Our conf.
EN	Gen	0.566	0.619	0.619	0.735
	Age	0.614	0.617	0.605	0.692
	Ext	0.185	0.182	0.181	0.165
	Neu	0.243	0.226	0.226	0.208
	Con	0.167	0.158	0.158	0.146
	Agr	0.173	0.183	0.183	0.162
	Ope	0.157	0.149	0.149	0.143
ES	Gen	0.760	0.760	0.760	0.820
	Age	0.400	0.404	0.416	0.580
	Ext	0.185	0.177	0.176	0.156
	Neu	0.243	0.220	0.220	0.202
	Con	0.161	0.163	0.162	0.154
	Agr	0.162	0.169	0.169	0.157
	Ope	0.183	0.183	0.183	0.168
IT	Gen	0.632	0.705	0.737	0.853
	Ext	0.159	0.162	0.162	0.121
	Neu	0.202	0.215	0.215	0.170
	Con	0.126	0.135	0.136	0.113
	Agr	0.159	0.165	0.165	0.150
	Ope	0.186	0.178	0.177	0.102
NL	Gen	0.611	0.344	0.333	0.633
	Ext	0.131	0.140	0.139	0.105
	Neu	0.206	0.205	0.204	0.156
	Con	0.122	0.125	0.125	0.101
	Agr	0.163	0.161	0.162	0.130
	Ope	0.121	0.122	0.122	0.104

Table 5. Results of our configuration and the synthetic baselines. Accuracy is reported for Gen and Age, RMSE is reported for Ext, Neu, Con, Agr, and Ope.