What do your look-alikes say about you? Exploiting strong and weak similarities for author profiling.

Notebook for PAN at CLEF 2015

Piotr Przybyła and Paweł Teisseyre

Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warsaw, Poland
p.przybyla@phd.ipipan.waw.pl
teisseyrep@ipipan.waw.pl

Abstract We describe a two-step procedure for author profiling, which first exploits language similarities between users and then aims at discovering more complex dependencies for dissimilar users. The method is motivated by the fact that authors using very similar vocabulary are likely to have similar traits. We use both word-based and text-based features, as well as relying on previous research. The proposed approach gives successful results, especially for gender and age prediction. Moreover, we show the most useful features using relevance measures based on random forests.

1 Introduction

This paper outlines our approach to author profiling task at the 13th PAN evaluation lab on uncovering plagiarism, authorship, and social software misuse [8]. The goal is to analyse a collection of tweets (in English, Spanish, Dutch and Italian) and discover its author's gender, age and personality traits: extraversion, stability, agreeableness, conscientiousness and openness. Unfortunately, the available amount of training data is very small: from 34 users for Dutch to 152 for English. As it seems very unlikely to observe new significant dependencies in such sets, we have decided to generate features basing on a collection of lexicons obtained in previous works. What is more, we have observed that authors using very similar vocabulary (the *look-alikes*) tend to have identical traits. We exploit this fact by performing a two-step prediction procedure: classifying a new item starts by finding a close neighbour; a full prediction model is used only in case nothing close enough could be found.

2 Features

In our approach, two groups of features are used: word-based and text-based. The word-based features represent numbers of occurrences of lemmas obtained with multilanguage *TreeTagger* [10]. The text-based features, computed as global statistics of text, include the following:

- `length` – average tweet length (number of characters),

- `wordLength` – average word length,
- `urls` – average number of URLs per tweet[1],
- `hashtags` – number of hashtags,
- `citations` – number of citations (@username),
- `capitals` – fraction of capital letters,
- `exclamations` – number of exclamation marks,
- `questions` – number of question marks,
- `emoticonsPos` – number of positive emoticons (recognized by a regular expression: `" [:;]\S*[\)DpP\]\ *]"`),
- `emoticonsNeg` – number of negative emoticons (recognized by a regular expression: `" :\S*[\(/\\\ |C]"`),
- `repeatedLetters` – fraction of repeated letters,
- `repeatedMarks` – fraction of repeated exclamation and question marks,
- `numbers` – number of numerical expressions (recognized by a regular expression: `" \d+([\.,]\d+)*"`),
- `errors` – number of spelling errors (obtained using multi-language *LanguageTool*),
- `yuleK` – vocabulary size estimated using Yule's K [16].

To improve the predictions, we have also taken into account previous research on text-based prediction of sentiment, emotions, etc. by including the following lexical features:

- for all languages: `SSPositive`/`SSNegative` – positive/negative sentiment score of collection of tweets, using *SentiStrength* tool [13],
- for English:
  - `NRCEmotion_*` – numerical value of 10 emotion associations (averaged per word[2]), using *NRC Word-Emotion Association Lexicon* [4],
  - `NRCTwitterSentiment` – sentiment value, using *NRC Twitter Sentiment Lexicon* [2],
  - `NRCHashtagSentiment(140)` – sentiment value, using *NRC Hashtag Emotion Lexicon* and *Sentiment140* lexicon [2],
  - `LexiconAFINN` – sentiment value, using AFINN Lexicon [6],
  - `MRC_*` – features from the MRC base [15]: familiarity, concreteness, imagery, meaningfulness (two measures) and age of acquisition,
  - `WWBPLexAge` and `WWBPLexGender` – usage of age- and gender-dependent lexicons from *World Well-Being Project* (WWBP) [9],
  - `WWBPAll*` – correlations with author features: gender, age and personality using data from WWBP [11],
- for Spanish: `SELEmotion_*` – numerical value of one of 6 emotions (joy, anger, fear, disgust, surprise, sadness), using *Spanish Emotion Lexicon* [12],
- for Dutch: `NLEmotion_*` – numerical value of valence, arousal, dominance and age of acquisition, using lexicon [5],

In total, we have obtained 56 features. Unfortunately, a great deal of them provides information only in case of English texts.

[1] All subsequent numbers are also averaged per tweet, unless noted otherwise.

[2] All subsequent values are also averaged per word.

3 Prediction

To predict traits (gender, age and personality) of Twitter users we apply a simple two-step procedure. The idea is to start with exploring close similarities between writings, and then try to discover more complex dependencies. More specifically, to predict traits for a new user, we first find the most similar user in the training data. If the similarity is sufficiently close, we assign traits of the found user to the new user. Otherwise, we use an advanced classification model to predict the traits. This approach is motivated by the fact that among large number of tweets one can easily find messages written by the same user. Moreover, it may happen that one person sends tweets from different Twitter accounts. So-called multiple Twitter accounts, which allow to boost users' presence in web, are becoming more and more popular. Finally, a very similar vocabulary can be shared by certain groups of users, having also similar features.

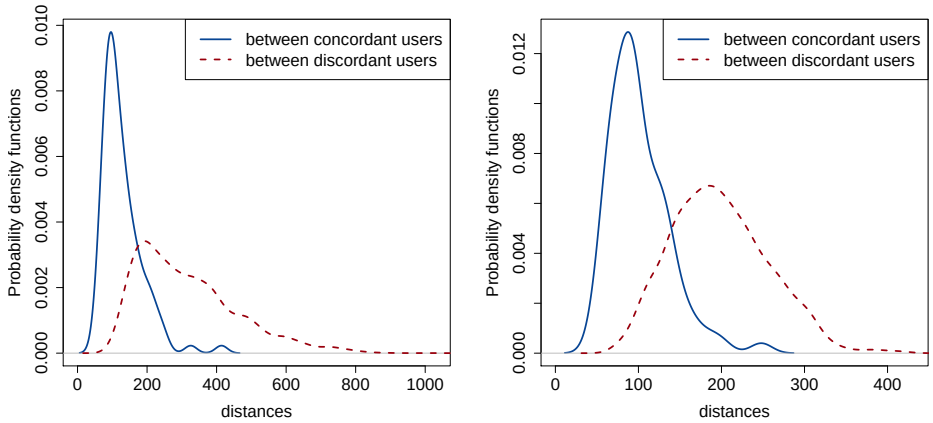


Figure 1. Smoothed histograms of distances between users for English (a) and Spanish (b).

Figure 1 shows normalised smoothed histograms of distances between concordant (having the same traits) and discordant users for English and Spanish. Although the histograms are partly overlapping, it is clear that distances among the first group are usually much smaller that in the second one. The advanced classification model used in the second step allows to discover more complex dependencies. The details of the whole procedure are given below.

Prediction Algorithm:

1. **Finding similar users in training data**. Here, we use two approaches, depending on the language of tweets.

- For English we build a classification model in which identifier of a group of concordant users (having the same traits) is used as a class variable. As a classification model, we use random forests [3], built on all available features. If the maximum of predicted probabilities for the new user is greater than a certain threshold p_{min}, we assign traits of a corresponding group to the new user.
- For other languages we simply find a nearest neighbour of the new user in the training data. To determine nearest neighbour, we use Euclidean distance and all available features. If the distance is less than a certain threshold d_{max}, we assign traits of the nearest neighbour to the new user.

2. **Prediction for dissimilar users**. If no similar users in training data are found, i.e. predicted probability of the best group is smaller that p_{min} (for English) or the distance to the nearest neighbour is greater than d_{max} (for other languages), we apply random forest method to predict each trait separately. We use all available features except word-based. For gender and age, decision trees are taken as base learners, whereas for personal traits regression trees are used. Other classification algorithms have also been tested (e.g. logistic regression) but they have yielded poorer results.

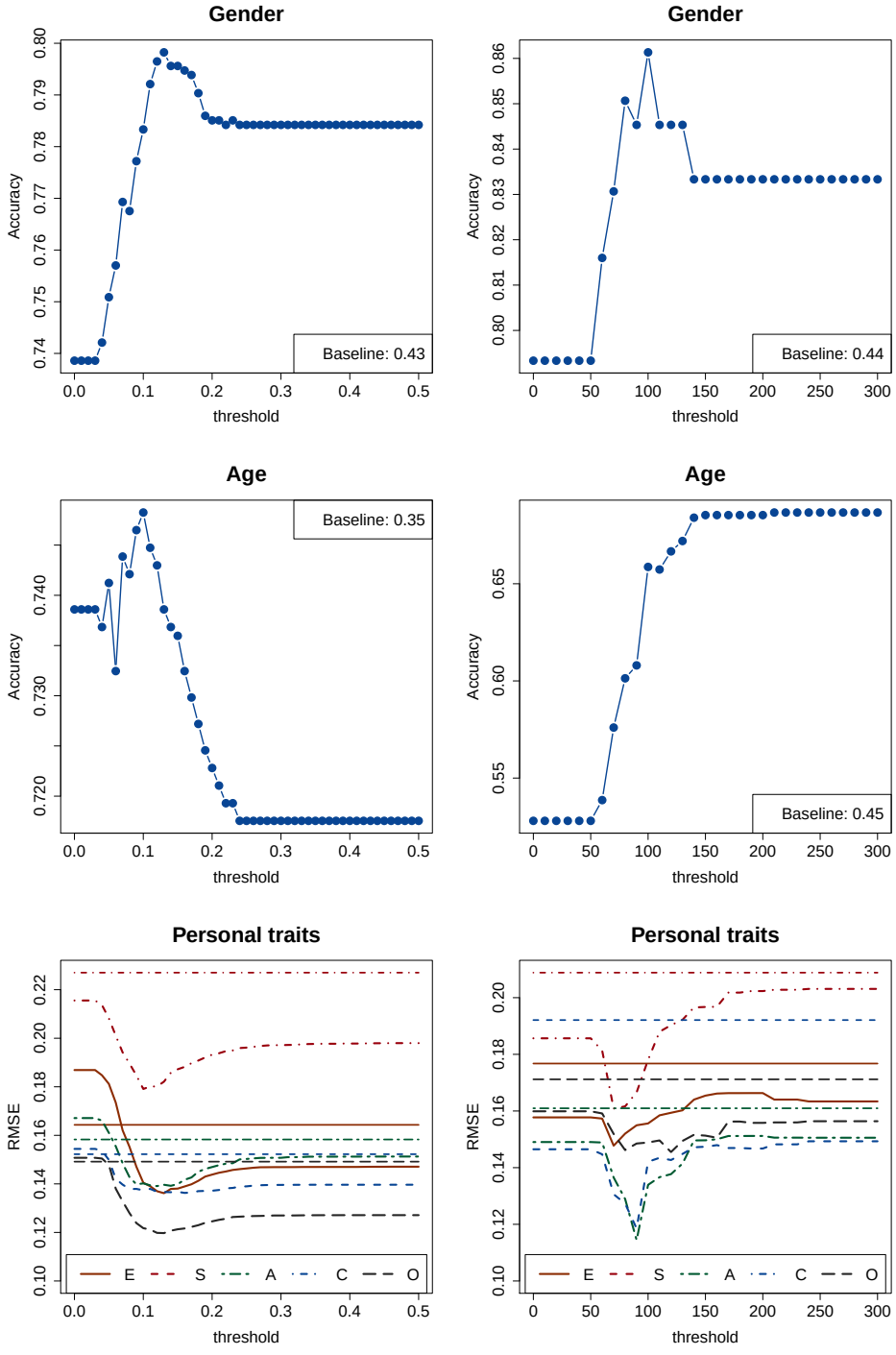
Observe that above procedure depends on the choice of threshold. If p_{min} is sufficiently small (for English) or d_{max} sufficiently large (for other languages), all users from training data are recognized as similar users and therefore only the first step of the above procedure is run. In the opposite case the full prediction model is always employed. To calibrate a threshold, we randomly split data (30 times) into training and testing parts and then compute averaged accuracy (gender and age) and mean error – RMSE (personal traits) for different values of threshold. Figure 2 shows the results for English and Spanish.

There is a clear optimum (maximum accuracy or minimum RMSE) for certain value of threshold. Note that for English the optimal value is common for all traits and equals $p_{min} \approx 0.12$. For Spanish an optimum is at $d_{max} \approx 90$ for gender and personal traits, whereas in case of age it is better to use nearest neighbour approach to all users. For the remaining languages we always apply nearest neighbour method (i.e. set $d_{max} = 0$), as the training sets are to small to build complex models.

4 Results

We have examined how the prediction procedure presented in Section 3 works with the set of features described in Section 2. As measures of performance we use accuracy (gender and age) and RMSE (personal traits). We randomly split data into training and testing parts in the following proportions: 75% for training and 25% for testing (for English and Spanish). For Italian and Dutch, due to small amount of data, we take only one observation for testing and the rest for training. The above procedure is repeated 30 times and the results are averaged over all runs. Classification procedure is implemented in `R` system [7] using libraries: `randomForest` [3], `FNN` [1] and `class` [14].

Results of our experiments (for an optimal value of threshold) are shown in Table 1. Numbers in brackets correspond to a baseline which is major class share (for classification) and mean value (for regression), calculated on training data. The third column



(a) (b)

Figure 2. Accuracy and RMSE with respect to threshold for gender, age and components of personality, for English (a) and Spanish (b). Horizontal lines correspond to baseline.

	Accuracy			RMSE					
	Gender	**Age**	**Gender&Age**	**E**	**S**	**A**	**C**	**O**	**Mean**
English	0.798	0.748	0.659	0.136	0.179	0.139	0.136	0.120	0.143
	(0.432)	(0.353)	(0.215)	(0.164)	(0.227)	(0.158)	(0.152)	(0.149)	(0.170)
Spanish	0.861	0.687	0.671	0.148	0.161	0.114	0.119	0.146	0.141
	(0.437)	(0.451)	(0.247)	(0.177)	(0.209)	(0.161)	(0.192)	(0.171)	(0.182)
Dutch	0.767	-	-	0.133	0.060	0.040	0.060	0.102	0.074
	(0.5)	-	-	(0.158)	(0.117)	(0.126)	(0.099)	(0.142)	(0.128)
Italian	0.900	-	-	0.071	0.043	0.036	0.023	0.029	0.031
	(0.5)	-	-	(0.121)	(0.199)	(0.097)	(0.085)	(0.117)	(0.124)

Table 1. Results of experiments; numbers in brackets correspond to baseline. Notation: E – Extraversion, S – Stability, A – Agreeableness, C – Conscientiousness, O – Openness.

includes joint accuracy for gender and age, whereas the last column contains RMSE, averaged over 5 personal traits. First, note that all the results exceed baseline. It is seen that gender and age identification are successful: we obtain accuracy 77%-90% for gender and 69%-75% for age. Moreover, simultaneous prediction of these two traits is also possible: the accuracy is about 3 times larger than the baseline. Personality assessment is a much more challenging task. Our experiments indicate that it is difficult to obtain an error significantly below the baseline.

Finally, we assess predictive power of the features using variable importance measure based on random forests. The measure pertains to average decrease of node impurity (Gini impurity index for classification and residual sum of squares for regression). The average is taken over all splitting nodes and over all trees used to construct an ensemble classifier. The measure shows usefulness of a given feature for prediction when random forest is used as a prediction tool. Figure 3 shows top 20 features for prediction of selected traits for English. The plot clearly shows that features pertaining to words collected from *World Well-Being Project* (WWBP) are among the most useful for prediction. Moreover, it is interesting that simple style-based features like message length, numbers of exclamation marks or citations seem to be relevant in case of age identification.

5 Conclusions

In this study we present a two-stage procedure for author profiling, which first exploits language similarities between users and then aims to discover more complex dependencies. The method is motivated by the fact that authors using very similar language tend to have identical traits. Interestingly, it turns out that combination of these two steps usually outperforms using each step separately. Our approach is based both on sets of word-based and text-based features. While we obtain successful results for gender and age prediction, the personality identification seems to be much more challenging – the error is slightly below the baseline. The assessment based on random forests shows high relevance of features using lexica from previous works. The results of experiments show many possibilities for future work. In our method, separate classification models

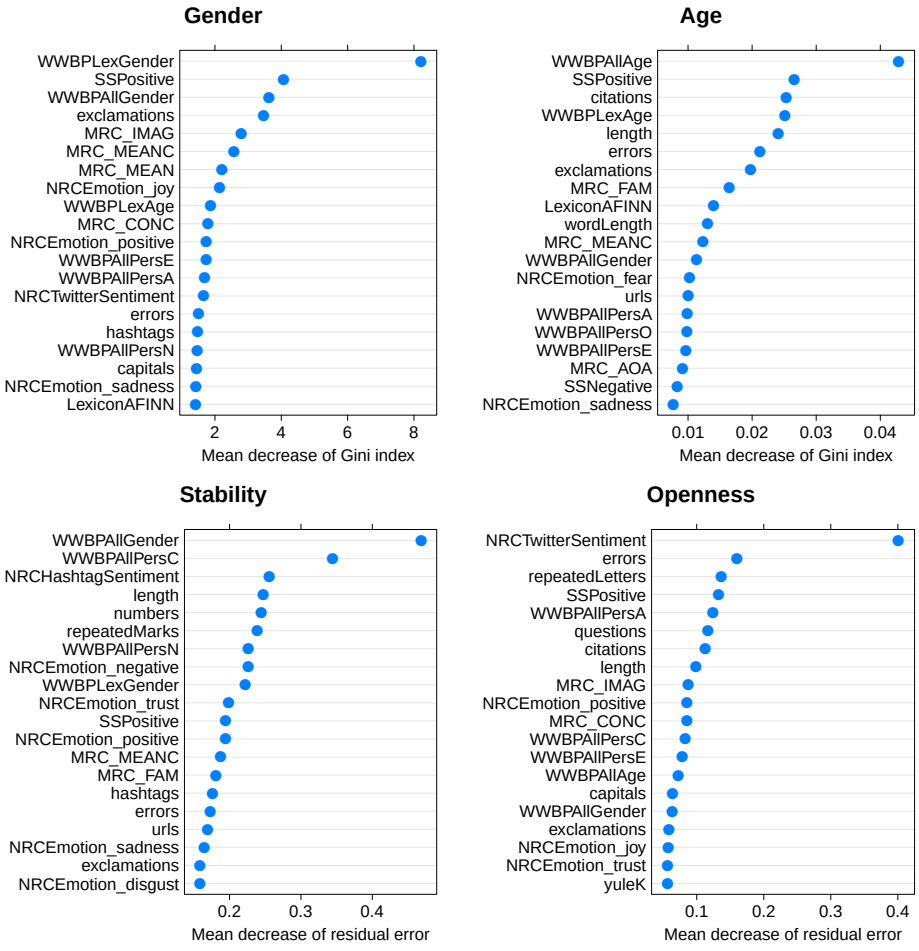


Figure 3. Feature importance measures based on random forest, for English.

are build for each trait – it is worthwhile to explore dependencies between the traits to improve the prediction performance. Secondly, in order to significantly improve personality identification, it seems necessary to look for new features. Finally, we believe that the advantages of using our two-stage procedure could be more clearly seen on larger corpus of tweets.

Acknowledgements

This study was supported by research fellowship within ”Information technologies: research and their interdisciplinary applications” agreement number POKL.04.01.01-00-051/10-00.

References

1. Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., Li, S.: FNN: Fast Nearest Neighbor Search Algorithms and Applications (manual) (2013)
2. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment Analysis of Short Informal Texts. Journal of Artificial Intelligence Research 50, 723–762 (2014)
3. Liaw, A., Wiener, M.: Classification and Regression by randomForest. R news 2, 18–22 (2002)
4. Mohammad, S.M., Turney, P.D.: Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence 29(3), 436–465 (2013)
5. Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., Van Harmelen, A.L., De Schryver, M., De Winne, J., Brysbaert, M.: Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. Behavior research methods 45(1), 169–77 (2013)
6. Nielsen, F.A.r.: A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In: Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages. vol. 718, pp. 93–98. CEUR-WS.org (2011)
7. R Core Team: R: A Language and Environment for Statistical Computing. Tech. rep., R Foundation for Statistical Computing (2013)
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. Sap, M., Park, G., Eichstaedt, J.C., Kern, M.L., Stillwell, D.J., Kosinski, M., Ungar, L.H., Schwartz, H.A.: Developing Age and Gender Predictive Lexica over Social Media. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1146–1151. Association for Computational Linguistics (2014)
10. Schmid, H.: Improvements In Part-of-Speech Tagging With an Application To German. In: Proceedings of the ACL SIGDAT-Workshop. pp. 47—-50. Association for Computational Linguistics (1995)
11. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.P., Ungar, L.H.: Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLOS ONE 8(9) (2013)
12. Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., Gordon, J.: Empirical study of machine learning based approach for opinion mining in tweets. In: Proceedings of the 11th Mexican international conference on Advances in Artificial Intelligence (MICAI'12). Lecture Notes in Computer Science, Springer-Verlag (2013)
13. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D.: Sentiment Strength Detection in Short Informal Text. Journal of the American Society for Information Science 61(12), 2544–2558 (2010)
14. Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S. Springer-Verlag (2002)
15. Wilson, M.: MRC psycholinguistic database: Machine-usable dictionary, version 2.00. Behavior Research Methods, Instruments, & Computers 20(1), 6–10 (1988)
16. Yule, G.U.: The Statistical Study of Literary Vocabulary. Cambridge University Press (1944)