PAN 2017: Author Profiling - Gender and Language Variety Prediction

Notebook for PAN at CLEF 2017

Matej Martinc[1,2], Iza Škrjanec[2], Katja Zupan[1,2], and Senja Pollak[1]

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
matej.martinc@ijs.si,skrjanec.iza@gmail.com,
katja.zupan@ijs.si,senja.pollak@ijs.si

Abstract We present the results of gender and language variety identification performed on the tweet corpus prepared for the PAN 2017 Author profiling shared task. Our approach consists of tweet preprocessing, feature construction, feature weighting and classification model construction. We propose a Logistic regression classifier, where the main features are different types of character and word n-grams. Additional features include POS n-grams, emoji and document sentiment information, character flooding and language variety word lists. Our model achieved the best results on the Portuguese test set in both—gender and language variety—prediction tasks with the obtained accuracy of 0.8600 and 0.9838, respectively. The worst accuracy was achieved on the Arabic test set.

Keywords: author profiling, gender, language variety, Twitter

1 Introduction

Recent trends in natural language processing (NLP) have shown a great interest in learning about the demographics, psychological characteristics and (mental) health of a person based on the text she or he produced. This field, generally known as author profiling (AP), has various applications in marketing, security (forensics), research in social psychology, and medical diagnosis. A thriving subfield of AP is computational stylometry, which is concerned with how the content and genre of a document contribute to its style [4].

One of the commonly addressed tasks in AP is the prediction of an author's gender, but other tasks include the prediction of language variety, age, native language, personality, region of origin or mental health of an author. Within this lively AP community, a series of scientific events and shared tasks on digital text forensic called PAN (Uncovering Plagiarism, Authorship, and Social Software Misuse)[3] have been organized. The first PAN event took place in 2011, while the first AP shared task was organized in 2013 [18].

[3] http://pan.webis.de/

In this paper, we describe our approach to the shared task of PAN AP for 2017 [20], which involves the construction of a model for gender and language variety identification of Twitter users. The rest of the paper is structured as follows: in Section 2 the findings from related work are presented. Section 3 describes the corpus and how it was preprocessed. In Section 4 we present the methodology, while Section 5 presents the results. In Section 6, we conclude the paper and present ideas for future work.

2 Related work

The earliest attempts in author profiling cover gender identification, starting with [8], who used parts of the British National Corpus. Other genres include literary texts [1], scientific papers [27], and emails [15].

The focus of AP has settled much on the social media, including languages other than English. Besides age and gender identification, the PAN shared task has addressed the prediction of personality type [16], setting the task into a cross-lingual [17,16,21,19] and cross-genre [17,21] environment. Since this year the corpus does not contain Dutch tweets, we only describe the findings of PAN AP 2016 task winners for English and Spanish. The goal was to built an age and gender classifier, whereby the model was trained on tweets and tested on blogs without the contestants knowing in advance the genre of the test set. The performance of contestants was evaluated by observing the classification accuracy for gender and age separately, and additionally taking into account the joint identification of both dimensions.

The team achieving the best score for gender classification (0.7564) in English was [11], who used the following features: word uni- and bigrams, character tetragrams, and the average spelling error, and logistic regression for learning. For gender classification in Spanish, the best result was obtained by Deneva [21], who achieved 0.7321 accuracy; a description of the system was not provided. For some contestants, the second order representation has proven useful. This was also the case with the overall winners for English [3], who trained a SVM model with RBF kernel and a SVM model with a linear kernel for age and gender, respectively. Their feature set comprised of unigrams and trigrams, employing also second order attributes and achieving a joint accuracy of 0.3974. The team [28] were the overall winners of the competition. Their linear SVM model performed with the overall accuracy of 0.5258 by employing a variety of features: word, character and POS n-grams, capitalization (of words and sentences), punctuation (final and per sentence), word and text length, vocabulary richness and hapax legomena, emoticons and topic-related words.

Language variety identification is a task of classifying different varieties of the same language by determining lexical and semantic variations between them [6]. Several studies performed classification on newspaper corpora, e.g. in Portuguese [30] and Spanish [32]. Data from social media is another popular resource for this task, e.g. in Spanish Twitter messages [10] or online comments in Arabic [25,29]. For the classification based on language variety several types of features have been considered. Lexical variation is explored with character and/or word n-grams [30,10,31,29,25], grammatical characteristics and syntax are represented in POS n-grams or their distribution [32,25,9]. Variation in orthography was used as a feature by employing a list of spelling variants [9]. Not only linguistic, but also historical and cultural differences were examined in [23] by observing the share of loan words in Brazilian and European Portuguese, while [24] used a so called ’black list’ of terms unwanted in Serbian, but accepted and used in Croatian.

3 Data set description and preprocessing

PAN 2017 training set consists of tweets in four different languages grouped by tweet authors, who are labeled by gender and language variety (Table 1). The number of authors for both categories (gender and variety) is balanced in every language. This training set was used for feature engineering, parameter tuning and training of the classification model.

Table 1. PAN 2017 training set structure

Language	Varieties	Authors	Tweets
English	Canada, Ireland, United States, Australia, New Zealand, Great Britain	3,600	360,000
Spanish	Argentina, Colombia, Venezuela, Spain, Chile, Mexico, Peru	4,200	419,998
Portuguese	Brazil, Portugal	1,200	120,000
Arabic	Egypt, Maghrebi, Gulf, Levantine	2,400	240,000

The following preprocessing steps were performed:

- *nonsense tweet removal*: on the English data set we discarded all tweets in which more than 90% of all tokens contain mistakes detected by a spell checker [7];
- *text reversal*: we reversed tweets in the Arabic data set since they are written from right-to-left.

Other preprocessing steps depend on feature construction and three data set transformations can be considered:

- *Tweets-cleaned*: replacing all hashtags, mentions and URLs with specific placeholders #HASHTAG, @MENTION, HTTPURL, respectively. Tweets-cleaned is also POS tagged (we used Averaged perceptron tagger from NLTK library[2] trained on POS tagged corpora for different languages found in NLTK);
- *Tweets-no punctuation*: removing punctuation from Tweets-cleaned;
- *Tweets-no stopwords*: stopwords are removed from Tweets-no punctuation. This preprocessing step is not used on Arabic language (Tweets-no stopwords transformation in Arabic is therefore identical to Tweets-cleaned transformation).

Finally, all tweets belonging to the same author are concatenated and used as one document in further processing.

4 Feature construction and classification model

The usefulness of character n-grams in authorship profiling has been proven before [16,17,26], as they contain information on punctuation, morphology and the lexis [4]. The setting with word uni- and bigrams, and character tri- and tetragrams was applied for gender and personality identification in [26]. For this reason most of the features used in our model were different types of n-grams. We also used other features, such as POS-tag sequences and features that depend on the use of external resources (an emoji list and word lists). We performed several different parameter tuning experiments (either manually or using the Scikit-learn grid search[4] to find best values) to try to find the best feature combination and parameters. All features were normalized with MinMaxScaler from the Scikit-learn library [13].

4.1 Features

The following n-gram features were used in our final model:

- *word unigrams*: calculated on lower-cased Tweets-no stopwords, TF-IDF weighting (parameters: minimum document frequency = 10, maximum document frequency = 80%);
- *word bigrams*: calculated on lower-cased Tweets-no punctuation, TF-IDF weighting (parameters: minimum document frequency = 20, maximum document frequency = 50%);
- *word bound character tetragrams*: calculated on lower-cased Tweets-cleaned, TF-IDF weighting (parameters: minimum document frequency = 4, maximum document frequency = 80%);
- *punctuation trigrams* (the so-called beg-punct [22], in which the first character is punctuation but other characters are not): calculated on lower-cased Tweets-cleaned, TF-IDF weighting (parameters: minimum document frequency = 10%, maximum document frequency = 80%);
- *suffix character tetragrams* (the last four letters of every word that is at least four characters long [22]): calculated on lower-cased Tweets-cleaned, TF-IDF weighting (parameters: minimum document frequency = 10%, maximum document frequency = 80%).

Other features used in the experiments were calculated on Tweets-cleaned data set transformation:

- *POS trigrams*: sequences of three POS tags, TF-IDF weighting (parameters: minimum document frequency = 10%, maximum document frequency = 60%);
- *emoji counts*: the number of emojis in the document, counted by using the list of emojis created by [12][5];

[4] http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[5] http://kt.ijs.si/data/Emoji_sentiment_ranking/

- *document sentiment*: the above-mentioned emoji list also contains the sentiment of a specific emoji, which allowed us to calculate the sentiment of the entire document by simply adding the sentiment of all the emojis in the document (as it turns out, this feature works better without normalizing the resulting sentiment with the number of all emojis in the document);
- *character flood counts*: we counted the number of times that three or more identical character sequences appear in the document;
- *language variety specific word lists*: according to [14] there are words that are specific for different language varieties. We managed to find an English spell checker dictionary[6] containing three different word list for three different English language varieties (United States, Canada and Australia). We calculated the intersection of these three word lists and removed the resulting common words from all three lists. In this way we obtained three language variety specific word lists, which enabled us to count the number of words appearing in a specific language variety list in every document. These features were only used in the English variety classification task.

We also experimented with TruncatedSVD topic modelling and Word2Vec embeddings but these features failed to improve the performance of our model so they were not included in the final model. Many different word count features (e.g., how many times a specific type of word appears in the document), punctuation count features and statistical features such as document length and average word length were also tested. All of these features were evaluated with chi2 feature selection utility from Scikit-learn[7] and proved statistically insignificant in relation to gender and variety target values. Moreover, they did not improve the performance of the model in the 10-fold cross-validation experiments on the training set, which is why they are not included in the final model.

4.2 Classification model

We tested several classifiers and different parameter sets. The following classifiers from Scikit-learn were tested:

- Linear SVM
- Logistic regression
- Random forest
- XGBoost (Extreme gradient boosting)

We also tested some classifier combinations:

- Logistic regression bagging
- Voting classifier with majority vote between Logistic regression, linear SVM and Random forest

Best results were obtained with Logistic regression. Bagging and voting did not improve the results. Logistic regression gave best results with C=1e2 and fit_intercept= False

[6] http://wordlist.aspell.net/dicts/

[7] http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html

parameters. With the help of Scikit-learn FeatureUnion[8], we were also able to specify the weights for different types of features we used. The weights were adjusted with the help of the following procedure:

1. Initialize all feature weights to 1.0.
2. Iterate the list of features. For every feature repeat adding or subtracting 0.1 to the weight until the accuracy of a 10-fold cross-validation is improving. When the best weight is found, move to the next feature on the list.
3. Repeat step 2 until the accuracy cannot be improved anymore.

The weights in our final Logistic regression model were the following:

- word unigrams and word bound character tetragrams: 0.8
- suffix character tetragrams: 0.4
- emoji and character flood counts, document sentiment and language variety specific word lists: 0.3
- POS trigrams: 0.2
- word bigrams and punctuation trigrams: 0.1

We considered adjusting weights for every task and language separately but initial experiments showed that no significant gains in accuracy can be achieved by doing this. This weight configuration proved optimal for both classification tasks and all the languages, which gave us some indication that no significant overfitting was taking place. Our classification model was therefore almost identical for all the languages and both tasks, with the exception of using language variety specific word lists as features in the English language variety task and using no POS trigrams as features in Arabic language.

5 Results

We present the accuracy of our model on the 10-fold cross-validation test as well as the accuracy of the model on the PAN 2017 official test set. The results of a 10-fold cross-validation test are shown in Table 2. All classes are balanced, so for gender the majority classifier's accuracy is 0.50. For language variety, the majority classifier would achieve 0.25 for Arabic, 0.50 for Portuguese, 0.143 for Spanish and 0.167 for English. As can be seen, the model performs best on Portuguese, where it achieved 0.8441 accuracy for gender and 0.9883 for the language variety prediction. The model reaches the lowest gender classification accuracy on Spanish and the lowest language variety classification accuracy on Arabic.

Accuracy results from the PAN 2017 official test set are presented in Table 3. The official PAN 2017 evaluator also measures the accuracy of the model in terms of predicting gender and language variety together (i.e., how many out of all the documents were correctly classified by both the gender and language variety), which is a measurement that was not employed in the 10-fold cross-validation experiments. Again, the model reached the best results on Portuguese, where it achieved 0.8600 accuracy for

[8] http://scikit-learn.org/stable/modules/generated/sklearn.pipeline.FeatureUnion.html

Table 2. Accuracy results of 10-fold cross-validation

	Gender	Language Variety
Arabic	0.8137	0.8345
Portuguese	**0.8441**	**0.9883**
Spanish	0.8059	0.9461
English	0.8280	0.8663

gender, 0.9838 accuracy for language variety prediction and 0.8463 accuracy for both. The model had the worst results for joint gender and language variety prediction on Arabic.

Table 3. Accuracy results on the official PAN 2017 AP test set

	Gender	Language Variety	Both
Arabic	0.8031	0.8288	0.6825
Portuguese	**0.8600**	**0.9838**	**0.8463**
Spanish	0.8193	0.9525	0.7850
English	0.8071	0.8688	0.7042

If we compare results from the 10-fold cross validation experiment and results from the official PAN 2017 test set, we can see that there are some differences. Surprisingly, Arabic is the only language where the results of the 10-fold cross-validation are better than the results on the official PAN 2017 test set on both classification tasks. On the contrary, the model achieved higher accuracy in both of these tasks on the Spanish official PAN 2017 test set. When it comes to English, higher accuracy for gender classification was achieved on the 10-fold cross-validation test and higher accuracy for language variety classification was reached on the official PAN 2017 test set (although the difference in accuracy is very small in this case). For Portuguese, higher accuracy for gender classification was achieved on the official PAN 2017 test set, while higher accuracy for language variety classification was obtained in the 10-fold cross-validation setting.

In general, we can conclude that differences in the accuracy measured on 10-fold cross-validation and official PAN 2017 test sets are not that large, meaning that our model did not overfit in most of the tasks in all the languages. The biggest difference in accuracy measurements is in English gender classification, where 10-fold cross-validation accuracy is more than 2% higher than on the official PAN 2017 test set. This suggests that some overfitting might have occurred in this case.

6 Conclusion and future work

In this paper we have presented our approach to the PAN 2017 author profiling task. We presented findings from the related work that were taken into consideration during the planning phase of our approach. We have also described the preprocessing techniques used, the methodology of our approach and the conducted experiments. Finally, we have presented the results achieved in the 10-fold cross-validation setting and on the official PAN 2017 test set. Our best results for the gender and language variety classification tasks in terms of accuracy were achieved for the Portuguese language and stand at 0.8600 and 0.9838, respectively. If we compare our performance with the results of other participants of PAN 2017, we were placed second in terms of joint accuracy achieved on both tasks, second in gender classification and third in language variety classification. Our model won on the task of gender classification in Arabic.

In our experience, the most difficult part of the task was finding the right features and properly weighting their combination. Our approach confirms the results from related work [22] that determined character n-grams as the most successful features in the AP tasks. Other n-grams, such as word unigrams and bigrams, also work well. The remaining features we used, i.e. POS tag sequences, emoji counts, character flood counts, language variety specific word lists and document sentiments, do not substantially contribute to the classification model accuracy but do, however, offer some new information to the classifier, so they can be considered as useful when combined with other features.

In the future, we plan to evaluate the model on different data sets to test and try to improve the cross-genre performance of the model. We will also consider a deep learning approach to gender and language variety classification. We also plan to address the gender classification task for other languages, such as Slovenian (there is a data set of Slovenian tweets and blogs with labeled gender [5]), Croatian and Serbian. We will also test our language variety classification model on the task of distinguishing between very similar languages, such as Serbian and Croatian.

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research Agency core research programme Knowledge Technologies (P2-0103) and the project “Resources, Tools and Methods for the Research of Nonstandard Internet Slovene" (J6-6842).

References

1. Argamon, S., Goulain, J.B., Horton, R., Olsen, M.: Vive la différence! text mining gender difference in french literature. Digital Humanities Quarterly 3(2) (2009)
2. Bird, S.: Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive presentation sessions. pp. 69–72. Association for Computational Linguistics (2006)
3. Bougiatiotis, K., Krithara, A.: Author profiling using complementary second order attributes and stylometric features. CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers (2016)
4. Daelemans, W.: Explanation in computational stylometry. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 451–462. Springer (2013)
5. Fišer, D., Erjavec, T., Ljubešić, N.: Janes v0.4: Korpus slovenskih spletnih uporabniških vsebin. Slovenščina 2.0 4(2), 67–99 (2016)

6. Franco-Salvador, M., Rangel, F., Rosso, P., Taulé, M., Martít, M.A.: Language variety identification using distributed representations of words and documents. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Lecture Notes in Computer Science. pp. 28–40. Springer International Publishing Switzerland (2015)
7. Kelly, R.: Pyenchant: A spellchecking library for python. http://pythonhosted.org/pyenchant/api/enchant.html, [Online; accessed 15-January-2017]
8. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
9. Lui, M., Cook, P.: Classifying English documents by national dialect. In: Proceedings of Australasian Language Technology Association Workshop. pp. 5–15. Association for Computational Linguistics (2013)
10. Maier, W., Gómez-Rodríguez, C.: Language variety identification in Spanish tweets. In: Language Technology for Closely Related Languages and Language Variants. pp. 25–35. Association for Computational Linguistics (2014)
11. Modaresi, P., Liebeck, M., Conrad, S.: Exploring the effects of cross-genre machine learning for author profiling in PAN 2016. CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers (2016)
12. Novak, P.K., Smailović, J., Sluban, B., Mozetič, I.: Sentiment of emojis. PloS one 10(12), e0144296 (2015)
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 12, 2825–2830 (2011)
14. Porta, J., Sancho, J.L.: Using maximum entropy models to discriminate between similar languages and varieties. In: Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects. pp. 120–128 (2014)
15. Prabhakaran, V., Reid, E.E., Rambow, O.: Gender and power: How gender and gender environment affect manifestations of power. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1965–1976. ACL (2014)
16. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at PAN 2015. In: CLEF 2015 Working Notes. CEUR (2015)
17. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the author profiling task at PAN 2014. In: CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers. CEUR (2014)
18. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. Notebook Papers of CLEF pp. 23–26 (2013)
19. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In: CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers. CEUR (2013)
20. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2017)
21. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In: CLEF 2016 Working Notes. CEUR-WS.org (2016)
22. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015. pp. 93–102 (2015), http://aclweb.org/anthology/N/N15/N15-1010.pdf

23. Soares da Silva, A.: Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese: endo/exogeneousness and foreign and normative influence. In: Advances in Cognitive Sociolinguistics. p. 41–84. De Gruyter Mouton (2010)
24. Tiedemann, J., Ljubešić, N.: Efficient discrimination between closely related languages. In: Proceedings of COLING 2012. p. 2619–2634. COLING (2012)
25. Tillmann, C., Al-Onaizan, Y., Mansour, S.: Improved sentence-level arabic dialect classification. In: Proceedings of the VarDial Workshop. pp. 110–119. Association for Computational Linguistics (2014)
26. Verhoeven, B., Daelemans, W., Plank, B.: Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In: 10th International Conference on Language Resources and Evaluation (LREC 2016) (2016)
27. Vogel, A., Jurafsky, D.: He said, she said: Gender in the acl anthology. In: Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries. pp. 33–41. ACL (2012)
28. Busger op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., Nissim, M.: Gronup: Groningen user profiling notebook for PAN at clef 2016. CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers (2016)
29. Zaidan, O.F., Callison-Burch, C.: Arabic dialect identification. Computational Linguistics 40(1), 171–202 (2014)
30. Zampieri, M., Gebre, B.G.: Automatic identification of language varieties: The case of Portuguese. In: Jancsary, J. (ed.) Proceedings of KONVENS 2012. pp. 233–237. ÖGAI (September 2012)
31. Zampieri, M., Gebre, B.G.: Varclass: An open source language identification tool for language varieties. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation. pp. 3305–3308. LREC (2014)
32. Zampieri, M., Gebre, B.G., Diwersy, S.: N-gram language models and POS distribution for the identification of Spanish varieties. In: Proceedings of TALN 2013. p. 580–587 (2013)