Including Dialects and Language Varieties in Author Profiling

Notebook for PAN at CLEF 2017

Alina Maria Ciobanu[1], Marcos Zampieri[2], Shervin Malmasi[3], Liviu P. Dinu[1]

[1]University of Bucharest, Romania
[2]University of Cologne, Germany
[3]Harvard Medical School, USA
alina.ciobanu@my.fmi.unibuc.ro

Abstract This paper presents a computational approach to author profiling taking gender and language variety into account. We apply an ensemble system with the output of multiple linear SVM classifiers trained on character and word n-grams. We evaluate the system using the dataset provided by the organizers of the 2017 PAN lab on author profiling. Our approach achieved 75% average accuracy on gender identification on tweets written in four languages and 97% accuracy on language variety identification for Portuguese.

1 Introduction

With vast amounts of texts available on social media, author (or authorship) profiling has become a popular research area in NLP. A number of characteristics such as age [19], gender [20], and native language [7,12] can be predicted based on the topics and the linguistic properties present in a person's writings.

The PAN labs[1] at CLEF have been providing a forum for scholars to evaluate authorship profiling approaches on user-generated content. Author profiling tasks organized in the past PAN labs included age, gender, and personality traits prediction [25,26]. This year, for the first time PAN includes language varieties and dialects from four languages: Arabic, English, Portuguese, and Spanish along with gender identification.[2]

This paper describes computational methods for gender and language variety identification on social media. Our approach builds on the experience acquired in the previous gender identification tasks of the PAN labs and the four editions of the Discriminating between Similar Languages (DSL)[3] shared task organized at the workshop on Similar

---

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

[2] In this paper we make a terminological distinction between (standard national) language varieties and dialects. We consider English, Spanish, and Portuguese to be pluricentric languages each of them including their own standard national language varieties. The situation of Arabic is, however, different as Modern Standard Arabic (MSA) co-exists with several Arabic dialects in a diglossic situation. Nevertheless, the challenges faced by systems trained to discriminate between similar languages, language varieties, and dialects are identical.

[3] http://ttg.uni-saarland.de/vardial2017/sharedtask2017.html

Languages, Varieties and Dialects (VarDial) [36,37,18,34]. The DSL shared tasks included all languages[4] and most of the dialects and language varieties included in the PAN lab 2017 thus establishing benchmarks for language variety and dialect identification.

2 Related Work

The inclusion of language varieties at PAN is motivated by the growing interest in dialect and language variety identification evidenced by several research papers and the aforementioned DSL and ADI shared tasks. Examples of such studies include Portuguese varieties [33,35,4], English varieties [11], Romanian dialects [6], Chinese varieties [31], and a number of studies on Arabic dialect identification [29,32,27,15].

The DSL and ADI shared task reports and their respective system description papers provide valuable information about successful approaches for dialect, language variety, and similar language identification. Successful approaches such as those by Goutte et al. (2014) [8], Malmasi and Dras (2015) [13], Malmasi and Zampieri (2016) [16], and Bestgen (2017) [2] rely on the combination of higher-order character n-grams (4 and above), word n-grams, POS tags in [3], and multiple linear classifiers such as SVMs and Naive Bayes arranged in ensembles and/or trained in a two-stage approach, in which first the language is identified and subsequently individual classifiers are trained to discriminate between language varieties or dialects of the same language.[5] An exception is the approach proposed by Ionescu and Butnaru (2017) [10] which achieved great results for Arabic dialect identification relying on kernel learning.

The main difference between the language variety sub-task at PAN and the DSL and ADI shared tasks is the kind of data provided by the organizers. The PAN challenge provides data collected from social media, whereas the data used in the DSL task comes from newspapers [28] and the ADI shared tasks used transcripts from broadcast speeches along with audio features [1]. With respect to the data, the most similar task to the PAN challenge is the 2014 TweetLID shared task [38] which included microblog posts from the languages spoken in the Iberian Peninsula and English.

3 Methods

3.1 Task and Data

The organizers of the PAN challenge on author profiling provided participants with a training set containing ~1,140,000 microblog posts from *Twitter*. Each post in the training set was annotated with the user's metadata including the language, language variety or dialect, and gender. A test set including unlabeled posts was released a month later.

---

[4] Arabic dialect identification (ADI) was a sub-task of the DSL 2016 and an individual task in the more comprehensive VarDial evaluation campaign 2017.

[5] Goutte et al. (2016) [9] provides a comprehensive evaluation of the first two editions of the DSL shared task.

The four languages and their respective varieties and dialects included in the PAN 2017 dataset are listed next.

- Arabic: Egypt, Gulf, Levantine, Maghrebi.
- English: Australia, Canada, Great Britain, Ireland, New Zealand, United States.
- Portuguese: Brazil, Portugal.
- Spanish: Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela.

The training set was annotated in XML format. Next we present an example of the meta-data for a male English speaker from the United States.

```
<author id="author-id" lang="en" variety="united states"
gender="male">
```

With the data provided by the PAN 2017 organizers in hand we trained SVM classifiers to identify both the gender and the language variety or dialect of users. Participants could choose to participate in any or both sub-tasks and we decided to participate in both.

Finally, it is worth noting that, unlike most NLP shared tasks, PAN requires participants to run their scripts in a virtual machine provided by the organizers. This ensures that all teams have the same computing power to participate in the challenge allowing full reproducibility [22].[6]

3.2 Approach

We use a single-label multi-class classification approach based on SVM ensembles, following the methodology proposed by Malmasi and Dras [13].

Classification ensembles are systems that combine the results of multiple classifiers, with the purpose of improving the overall performance. Ensembles have been successfully used in various research areas, such as complex word identification [14] or grammatical error diagnosis [30]. The individual classifiers can differ in various regards, such as training data, features or classification methods.

In our system, the classifiers differ in terms of features. We use character n-grams (with n in $\{1, ..., 6\}$) and word n-grams (with n in $\{1, 2\}$) and build a classifier for each type of feature. Thus, our ensemble consists of eight individual classifiers. To combine the classifiers, we employ a fusion method based on the probability estimates provided by the individual classifiers: the predicted probabilities for each class are added, and the prediction of the ensemble is the class with the highest sum. We use the SVM implementation provided by Scikit-learn [21], based on the Libsvm library [5].

We train the ensembles individually for predicting gender and language varieties. We perform 3-fold cross-validation on the training dataset for hyperparameter tuning, for each classifier, searching for the optimal value of C in $\{10^{-5}, ..., 10^5\}$.

---

[6] The PAN labs use TIRA (http://www.tira.io/) for reproducibility.

4 Results

In the next two sections we present the results obtained by our method. Section 4.1 presents the results obtained using cross-validation on the training set. Section 4.2 presents the official results obtained using the PAN author profiling test set released by the shared task organizers over a month after the training set was released.

4.1 Cross-Validation

The cross-validation results are reported in Table 1 with the best results presented in bold. We note that the highest joint accuracy (when both the gender and language variety are correctly predicted together) is obtained for Portuguese, where the system obtains 0.75 accuracy. For gender identification, the highest accuracy of 0.79 is obtained for English, while language variety is best predicted for Portuguese, with 0.97 accuracy. Portuguese also obtains the highest average accuracy of 0.83 (average of gender, language variety and joint accuracy).

Table 1. Cross-validation accuracy on the training set for gender and language variety.

Language	Gender	Variety	Joint	Average
Arabic	0.73	0.75	0.57	0.68
English	**0.79**	0.75	0.59	0.71
Portuguese	0.77	**0.97**	**0.75**	**0.83**
Spanish	0.71	0.90	0.64	0.75

The high results obtained for Portuguese were not surprising, as there were only two Portuguese varieties in the dataset, from Brazil and from Portugal. The dataset included more varieties and dialects from the other four languages, namely: six English varieties, seven Spanish varieties, and four Arabic dialects.

The individual classifiers do not outperform, in any case, the ensembles. Portuguese is the only language for which the best individual performance equals the performance of the ensembles. For the others, the improvement reaches a maximum of 0.08 in accuracy (for the English joint prediction) when using ensembles. For three languages out of four (English, Spanish and Portuguese), word unigrams obtain the highest joint accuracy from all the individual classifiers. For Arabic, character 4-grams obtain the highest joint accuracy. As far as the language variety and gender labels are concerned, character 4-grams, character 5-grams and word unigrams obtain better results than the other types of features. For both gender and language variety identification, the best results are obtained for Portuguese, using character 4-grams for gender identification and word unigrams for language variety identification.

4.2   Test Set

In the official evaluation carried out on the test set by the PAN organizers our system was ranked 13th among 22 participating teams in both sub-tasks. The system achieved 0.7842 average average accuracy for language variety and gender identification. The results and ranks are described in more detail in the PAN labs report [23] and in the author profiling task report [24].

In Table 2 we present the results obtained for language variety identification. For reference we provide two baselines provided by the organizers: the BOW-baseline, a bag-of-words model with the 1,000 most frequent items and the STAT-baseline, a simple majority class baseline. As observed in the cross-validation experiments, the best results in the test set were also obtained when discriminating between the two Portuguese varieties achieving 0.9788 accuracy. On language variety identification our system achieved an average performance of 0.8524 accuracy ranking 11th among 22 shared task entries.

Table 2. Test set accuracy results for language variety identification.

Rank	Arabic	English	Portuguese	Spanish	Average
11th of 22	0.7569	0.7746	0.9788	0.8993	0.8524
BOW-baseline	0.3394	0.6592	0.9712	0.7929	0.6907
STAT-baseline	0.2500	0.1667	0.5000	0.1429	0.2649

In Table 3 we present the results obtained for gender identification with tweets from different languages along with the two aforementioned baselines. This is a binary classification setting in which the systems are trained to discriminate between tweets written by male and female writers. The variable gender was constant between all languages whereas the number of varieties and dialects for each language varied between 2 for Portuguese and 7 for Spanish. For this reason we observed that the results across languages for gender identification varied much less than the results obtained on language variety/dialect identification.

Table 3. Test set accuracy results for gender identification per language.

Rank	Arabic	English	Portuguese	Spanish	Average
12th of 22	0.7131	0.7642	0.7713	0.7529	0.7504
BOW-baseline	0.5300	0.7075	0.7812	0.6864	0.6763
STAT-baseline	0.5000	0.5000	0.5000	0.5000	0.5000

Our method obtained the best results for Portuguese tweets achieving 0.7713 and the lowest results for Arabic achieving 0.7131 accuracy. The average performance of our method on gender identification was 0.7504 accuracy ranking 12th among 22 shared task entries.

The results presented in this section indicate that our approach performed substantially better than the two baselines provided and it was consistently ranked in the middle of the table both for language variety and for gender identification. Even though the results obtained by our method were not low, taking the experience obtained in the past PAN labs and DSL shared tasks into account we expected our system to rank higher in the official scores table. Possible factors that may have influenced the performance of our method are: 1) the type of dataset used at PAN which contain very short and non-standard texts, 2) the large size of the dataset that might have made possible for the other teams to use innovative approaches (e.g. deep learning), and 3) our implementation of the classifier which might not have been optimal. A thorough analysis of the misclassified instances is being carried out to determine the reasons for this outcome and possible ways to improve our system's performance.

5 Conclusion

This paper presented an SVM ensemble-based system trained on character and word n-grams developed for author profiling tested on the PAN 2017 dataset which takes gender and language variety/dialect identification into account. The approach described in our submission was inspired by successful submissions to past editions of the PAN task on gender identification, to the Discriminating between Similar Languages (DSL), and to Arabic Dialect Identification (ADI) shared tasks, the last two organized at the VarDial workshop.

In the training set cross-validation stage, our best results for gender identification were obtained on English data, 0.79 accuracy, and the best results for language variety identification were obtained for Portuguese, 0.97 accuracy. In the official evaluation carried out on the test set our system was ranked 11th on language variety identification and 12th on gender identification out of 22 submissions achieving 0.85 and 0.75 accuracy respectively.

To the best of our knowledge, the PAN labs 2017 was the first shared task to include language varieties and dialects in author profiling opening avenues for future research. Regarding our system's performance, there is still room for improvement. We are currently investigating ways to improve our system's performance by testing a meta-classifier which achieved very good results on German dialect identification [17].

Acknowledgement

References

1. Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S.H., Glass, J., Bell, P., Renals, S.: Automatic Dialect Detection in Arabic Broadcast Speech. In: Proceedings of INTERSPEECH (2016)
2. Bestgen, Y.: Improving the Character Ngram Model for the DSL Task with BM25 Weighting and Less Frequently Used Feature Sets. In: Proceedings of the VarDial Workshop (2017)
3. Bestgen, Y.: Improving the character ngram model for the DSL task with BM25 weighting and less frequently used feature sets. In: Proceedings of the VarDial Workshop (2017)
4. Castro, D.W., Souza, E., Vitório, D., Santos, D., Oliveira, A.L.: Smoothed N-gram based Models for Tweet Language Identification: A Case Study of the Brazilian and European Portuguese National Varieties. Applied Soft Computing (2017)
5. Chang, C.C., Lin, C.J.: LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology 2(3), 27:1–27:27 (2011)
6. Ciobanu, A.M., Dinu, L.P.: A Computational Perspective on Romanian Dialects. In: Proceedings of LREC (2016)
7. Gebre, B.G., Zampieri, M., Wittenburg, P., Heskes, T.: Improving Native Language Identification with TF-IDF Weighting. In: Proceedings of the BEA workshop (2013)
8. Goutte, C., Léger, S., Carpuat, M.: The NRC System for Discriminating Similar Languages. In: Proceedings of the VarDial Workshop (2014)
9. Goutte, C., Léger, S., Malmasi, S., Zampieri, M.: Discriminating similar languages: Evaluations and explorations. In: Proceedings of LREC (2016)
10. Ionescu, R.T., Butnaru, A.: Learning to identify Arabic and German dialects using multiple kernels. In: Proceedings of the VarDial Workshop (2017)
11. Lui, M., Cook, P.: Classifying English Documents by National Dialect. In: Proceedings of ALTA (2013)
12. Malmasi, S., Cahill, A.: Measuring Feature Diversity in Native Language Identification. In: Proceedings of the BEA Workshop (2015)
13. Malmasi, S., Dras, M.: Language identification using classifier ensembles. In: Proceedings of the VarDial Workshop (2015)
14. Malmasi, S., Dras, M., Zampieri, M.: LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles. In: Proceedings of SemEval (2016)
15. Malmasi, S., Refaee, E., Dras, M.: Arabic Dialect Identification using a Parallel Multidialectal Corpus. In: Proceedings of PACLING (2015)
16. Malmasi, S., Zampieri, M.: Arabic Dialect Identification in Speech Transcripts. In: Proceedings of the VarDial Workshop (2016)
17. Malmasi, S., Zampieri, M.: German dialect identification in interview transcriptions. In: Proceedings of the VarDial Workshop (2017)
18. Malmasi, S., Zampieri, M., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J.: Discriminating between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task. In: Proceedings of the VarDial Workshop (2016)
19. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: "How Old Do You Think I Am?"; A Study of Language and Age in Twitter". In: Proceedings of ICWSM (2013)
20. Nguyen, D.P., Trieschnigg, R., Doğruöz, A., Gravel, R., Theune, M., Meder, T., de Jong, F.: Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment. In: Proceedings of COLING (2014)
21. 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 (2011)

22. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14) (2014)

23. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Initiative (CLEF 17) (2017)

24. 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: Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings (2017)

25. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: Proceedings of CLEF (2015)

26. 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. Proceedings of CLEF (2016)

27. Sadat, F., Kazemi, F., Farzindar, A.: Automatic Identification of Arabic Language Varieties and Dialects in Social Media. In: Proceedings of the SocialNLP Workshop (2014)

28. Tan, L., Zampieri, M., Ljubešić, N., Tiedemann, J.: Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection. In: Proceedings of the BUCC Workshop (2014)

29. Tillmann, C., Mansour, S., Al-Onaizan, Y.: Improved Sentence-Level Arabic Dialect Classification. In: Proceedings of the VarDial Workshop (2014)

30. Xiang, Y., Wang, X., Han, W., Hong, Q.: Chinese Grammatical Error Diagnosis Using Ensemble Learning. In: Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications. pp. 99–104 (2015)

31. Xu, F., Wang, M., Li, M.: Sentence-level dialects identification in the Greater China region. International Journal on Natural Language Computing (IJNLC) 5(6) (2016)

32. Zaidan, O.F., Callison-Burch, C.: Arabic Dialect Identification. Computational Linguistics 40(1), 171–202 (2014)

33. Zampieri, M., Gebre, B.G.: Automatic Identification of Language Varieties: The Case of Portuguese. In: Proceedings of KONVENS (2012)

34. Zampieri, M., Malmasi, S., Ljubešic, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., Aepli, N.: Findings of the VarDial Evaluation Campaign 2017. Proceedings of the VarDial Workshop (2017)

35. Zampieri, M., Malmasi, S., Sulea, O.M., Dinu, L.P.: A Computational Approach to the Study of Portuguese Newspapers Published in Macau. In: Proceedings of the NLP Meets Journalism Workshop (2016)

36. Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J.: A report on the DSL shared task 2014. In: Proceedings of the VarDial Workshop (2014)

37. Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J., Nakov, P.: Overview of the DSL shared task 2015. In: Proceedings of the LT4VarDial Workshop (2015)

38. Zubiaga, A., San Vicente, I., Gamallo, P., Pichel, J.R., Alegria, I., Aranberri, N., Ezeiza, A., Fresno, V.: Overview of TweetLID: Tweet language identification at SEPLN 2014. In: Proceedings of the TweetLID Workshop (2014)