Subword-based Deep Averaging Networks for Author Profiling in Social Media

Notebook for PAN at CLEF 2017

Marc Franco-Salvador, Nataliia Plotnikova, Neha Pawar, and Yassine Benajiba

Symanto Research, Nuremberg, Germany
{marc.franco,nataliia.plotnikova,
neha.pawar,yassine.benajiba}@symanto.net

Abstract Author profiling aims at identifying the authors' traits on the basis of their sociolect aspect, that is, how language is shared by them. This work describes the system submitted by Symanto Research for the PAN 2017 Author Profiling Shared Task. The current edition is focused on language variety and gender identification on Twitter. We address these tasks by exploiting the morphology and semantics of the words. For that purpose, we generate embeddings of the authors' text based on subword character n-grams. These representations are classified using deep averaging networks. Experimental results show competitive performance in the evaluated author profiling tasks.

1 Introduction

Author profiling aims at identifying the authors' traits on the basis of their sociolect aspect, that is, how language is shared by them. It is used to determine language variety, gender, age, and personality type, among others. This task is specially attractive to industry representatives and particularly helpful for author opinion segmentation in social media. For instance, identifying the geographical distribution and gender of opinion authors may help to improve marketing campaigns. The task is also important for digital text forensics. Given a threat, knowing the possible author traits may help to its identification.

The Uncovering Plagiarism, Authorship, and Social Software Misuse[1] (PAN) evaluation lab at the Conference and Labs of the Evaluation Forum[2] (CLEF) promotes research and innovation in digital text forensics. Its Author Profiling Shared Task set the objective of classifying authors' traits in several subtasks. These include the identification of age, cross-genre age, personality traits, and gender in social media. The current edition[3] focuses on language variety and gender identification on Twitter.

Both morphological [1,6] and semantic [7,2] features have proven to be highly discriminant in author profiling. To build on research, we exploit in this work word morphology and semantics to identify the authors' language variety and gender. We present

[1] http://pan.webis.de/
[2] http://www.clef-initiative.eu/
[3] http://pan.webis.de/clef17/pan17-web/author-profiling.html

an approach based on word embeddings which in turn are generated using the subword information, i.e., by means of character n-gram embeddings [3]. We classify the author traits using deep averaging networks, a recent technique which magnifies the most discriminant dimensions contained within an embedding average. This has been demonstrated to be a fast and competitive approach in several text classification tasks [10] — rivalling the recurrent or convolutional neural networks performance.

The rest of the work is structured as follows: in Section 2 we provide an overview of the state of the art in author profiling. In Section 3 we describe the system we employed for the PAN 2017 Author Profiling Shared Task. Next, in Section 4 we conduct our evaluation and discussion of the results. Finally, we draw our conclusions in Section 5.

2 Related Work

Authorship attribution [12], the task of identifying authors' stylistic discriminators, set the stage for the author profiling task. The use of stylistic features such as character and part-of-speech (PoS) n-grams, as well as spelling and grammatical errors, allowed us to identify authors' native language [13]. Similarly, [26] identified age and gender in blogs using stylistic and content word features. The popularity of author profiling motivated the organization of several workshops and shared tasks.

The Native Language Identification Shared Task [27] allowed participants to classify English essays representing eleven native languages. The Shared Task on Discriminating between Similar Languages (DSL) set the objective of classifying texts representing several sets of closely related languages and language varieties [29,30,17]. Since 2013, the PAN evaluation lab organized the Author Profiling Shared Task. The first two editions focused on age and gender identification [22,21]. In addition to these two tasks, personality traits recognition was included in 2015 [19]. Finally, the focus of the 2016 edition was cross-genre age and gender identification [24][23].

This year, the PAN author profiling track is focused on the tasks of language variety and gender identification. Regarding the latter, most of the recent work on gender identification originated in the PAN evaluation lab. The system winner of the 2013-2015 editions is based on a representation for documents which captures discriminative and subprofile-specific information [14]. Similar to the early work on the subject, the best performing system in 2016 employed content words, emoticons, and stylistic features [4].

The language variety identification task has attracted much interest in the last few years. Character n-grams and other features have been employed to identify varieties of Portuguese in news texts [28], Arabic in blogs and forums [25], and Spanish in tweets [15]. Word embeddings were used to classify varieties of Spanish from blogs and journalistic texts [7,8]. Also in the Spanish blogs domain, [20] a low dimensional model based on text statistics was employed. The best performing system of DSL 2015 [16] used an ensemble of models based on word and character n-grams.

Unlike the majority of author profiling researchers, which employ stylistic and lexical features, our approach is based on character n-gram word embeddings, with exploit the morphology and semantics of words. This choice has also been driven by our motivation to experiment with a pipeline that could be replicated fairly simply by researchers

who want to compare results and practitioners in need of a simple, yet accurate, pipeline to perform author profiling.

3 Proposed Approach

In this section we describe the system we designed for language variety and gender identification on Twitter. First, in Section 3.1 we describe our data preprocessing. Next, in Section 3.2 the embedding representations are described. Finally, in Section 3.3 we detail our classifier.

3.1 Preprocessing

We preprocess each text with tokenization, word lowercase, and removing URLs. We use the Tweet NLP[4] tokenizer, which is specific for English tweets. We slightly modified its regular expressions to consider Arabic, Portuguese, and Spanish punctuation, e.g. '¿' and '¡' were included for Spanish.

3.2 Subword Character n-gram Embeddings

In recent years, word embeddings replaced the bag-of-words (BOW) representation as the standard for text feature extraction.[5] These representations are low d-dimensional real-valued vectors which capture semantic and syntactic aspects of text. The continuous skip-gram model [18] of the word2vec toolkit is the preferred alternative to generate the embeddings.

We should note the importance of morphology in author profiling. For instance, the derivation of words is a discriminant feature in English language variety identification, e.g. *regulariz**ed*** vs. *regulari**sed***. As an additional example, the morphological refraction is indicative of gender in Latin languages, e.g. *profe**sor*** vs. *profe**sora*** in Spanish (male and female *professor* word translation, respectively).

In this work we use a recent variant of the continuous skip-gram model [3] which generates word embeddings exploiting the words' morphology by means of character n-gram embeddings. In addition to helping better capture the morphological nuances that we previously mentioned, a character based embedding model also helps to create robust classification models in the presence of typos and abbreviations as is usually the case in social media data.

When it comes to learning these embeddings, the main difference of this *subword* model is in the scoring function used to estimate the probability of observing a context word w_c given a target word w_t. The original model used the scalar product of the word vectors as scoring: $s(w_t, w_c) = u_{w_t}^T v_{w_c}$, where u_{w_t} and v_{w_c} are vectors in $\mathbb{R}^d$. The subword model uses instead a scoring function which represents the target word as the sum of its character n-gram vectors:

[4] `http://www.cs.cmu.edu/~ark/TweetNLP/`

[5] We note the increasing number of papers published at the ACL conference with "word embeddings" or "distributed representations" as part of the title: 0 (2013), 3 (2014), 15 (2015), and 29 (2016).

$$s(w, c) = \sum_{g \in \mathcal{G}_w} z_g^T v_c, \tag{1}$$

being $\mathcal{G}_w\{1, ..., G\}$ the set of n-grams of the word w, and z_g and v_c vectors in $\mathbb{R}^d$. Key of the model's design is the use of a hashing function to map n-grams to integers that represent the vector index. This makes the model memory efficient and provides with an additional feature: it does not produce out-of-vocabulary words. The embedding of an unknown word is created by extracting its n-grams and doing the average of the vectors with the indexes returned by the hash function. For more details about the model please refer to its original work.

We generate a word embedding inventory for the training partition (see Section 4.1) of each language using the FastText library.[6] We use 300-dimensional vectors, context windows of size 10, 20 negative words for each sample, 15 epochs, and 2M hashed character n-gram vectors. We extract n-grams with length in $[3, ..., 6]$. We post-process and enrich the embeddings with a proprietary model © Symanto Research.

3.3 Deep Averaging Networks

A standard method to obtain vector representations of text consists on computing the average of the word embeddings [5]. This embedding composition method obtained good results in language variety identification [7]. However, the longer the text, the more abstract the resulting embedding is.

In this work we classify using Deep Averaging Networks (DAN) [10]. As illustrated in Figure 1, this model receives as input the word embeddings of the text. First, a composition layer is put in place to average those embeddings. It proceeds then to use one or many non-linear hidden layers to transform the computed average. Finally, a softmax layer is used for prediction. The rationale behind DAN is that the non-linear transformations applied to the average allow to magnify and capture subtle variations in a more precise manner. As reported in the original paper, this approach can outperform syntactically informed approaches despite its simplicity.

Our hidden layers have size equal to the embedding one and use the rectified linear units (ReLU) [9] as activation function. We use the cross-entropy loss function. The number of hidden layers is determined in Section 4.2. We optimize the neural network weights with Adam [11], learning rate $= 0.001$ and 100 epochs, using the parameters indicated on its original work. We should note that our word embeddings are static so we do not allow the model to modify them.

4 Evaluation

In this section we evaluate our approach in the PAN 2017 Author Profiling Shared Task.

[6] `https://github.com/facebookresearch/fastText`

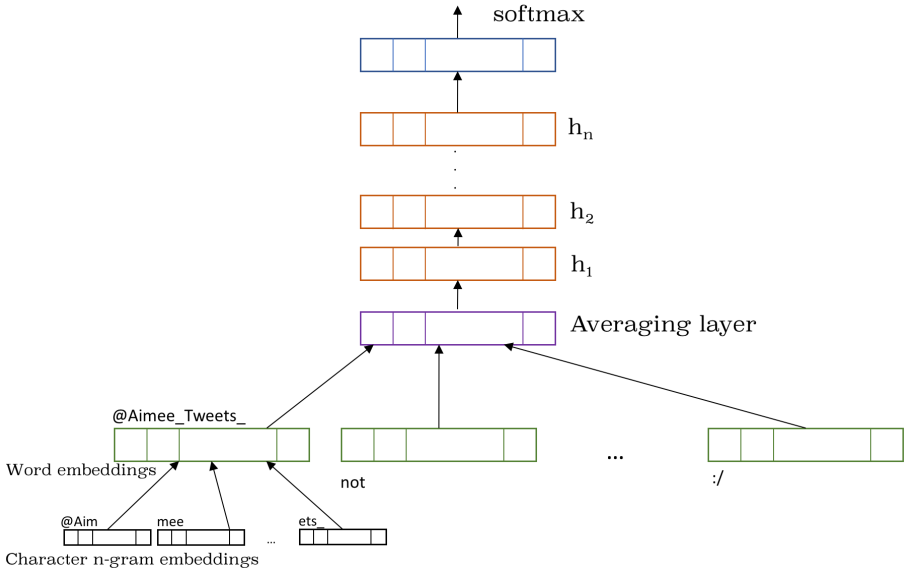


Figure 1: Illustration of the DAN architecture

4.1 Datasets and Tasks Setting

Dataset The objective of the PAN 2017 author profiling shared task is to identify the language variety and gender of Twitter users. Its corpus contains four languages and nineteen language varieties:

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

Next, we mention some key remarks about the dataset. The language of the user is known, so the dataset is composed by four partitions. In Table 1 we show the statistics. The labels are balanced at language variety and gender level. Finally, each Twitter user is represented by a set of approximately 100 tweets.[7]

In this work, we concatenate the user tweets to have an unique instance. We explored other alternatives, as the independent classification of the tweets with a subsequent sum of the class probabilities [7]. However, with this dataset, we obtained higher results after concatenating the tweets.

Methodology We compare our results with those obtained by the random baseline, a BOW model classified with random forest, a model based on continuous skip-gram embedding averages classified with logistic regression, and a model based on the subword

[7] Each tweet is composed by up to 140 characters.

Statistic	Arabic	English	Portuguese	Spanish
Training users	2,400	3,600	1,200	3,200
Test users	800	1,200	400	1,400
Language varieties	4	6	2	7

Table 1: Statistics of the PAN 2017 author profiling shared task dataset.

embedding (see Section 3.2) averages classified with logistic regression. In the rest of the evaluation we refer to these models as Random, BOW, skip-gram emb., and subword emb., respectively. The prototype of our model (henceforth simply referred to as DAN) was designed using 10-fold cross-validation over the training sets. The parameter selection uses the same setting. The official measure of the competition is the accuracy. The ranking of the shared task participants is estimated as follows: i) for each language, the PAN organizers calculate individual accuracies for gender and variety identification; ii) they calculate the accuracy when both variety and gender are properly predicted together; and iii) the final ranking is obtained by averaging those accuracy values obtained per language.

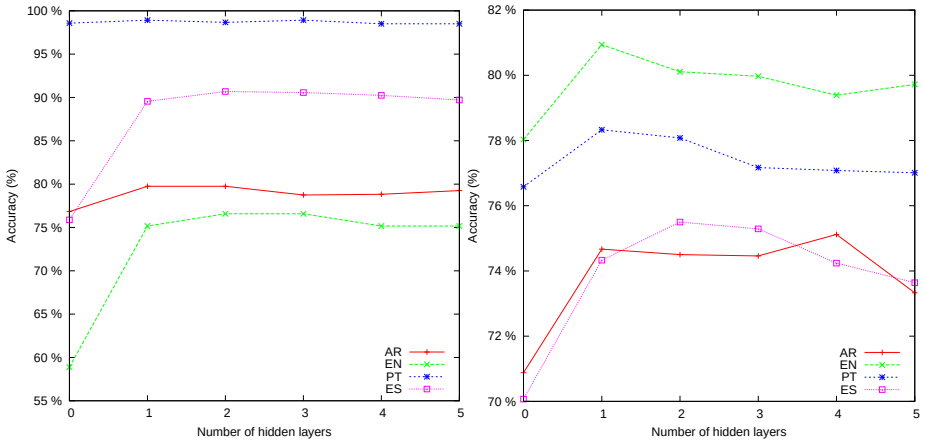
4.2 Parameter Selection

We noticed during our experimentation phase that the performance of DAN is very sensitive to the number of hidden layers, which differ in function of the task and dataset. In Figure 2 we show the accuracy depending on the number of hidden layers, task, and language. As you can see, the two tasks benefit from adding layers after composition/averaging one. The best performance for language variety identification is achieved using two layers. In contrast, the optimal number of hidden layers for gender identification differs depending on the language. We use the best parameters determined in this section for the rest of the evaluation.

4.3 Results and Discussion

In this section we compare and discuss the results of our system. In Table 2 we show the development experiments and the comparison with the baseline models (see Section 4.1) using 10-fold cross-validation over the training set. As we can see, the three embedding-based models outperform BOW, the only purely lexical approach. The continuous skip-gram embedding averages classified with logistic regression obtain better results than the subword embedding averages in tasks such as language variety identification in Arabic or gender in Portuguese. However, the latter model offers in average higher results than the skip-gram one. Finally, DAN, using the same subword embeddings, obtains the highest results and proves that deep averaging networks are useful in author profiling to magnify the most discriminant values contained in an embedding average.

In Table 3 we show the results using the official test set of the shared task. This table also includes the *joint* accuracy, which is employed by organizers to determine the best



(a) Language variety

(b) Gender

Figure 2: Deep averaging networks accuracy in function of its number of hidden layers.

system, i.e., when both variety and gender are properly predicted together. As we can see, DAN's results are in line with those obtained using the 10-fold cross-validation setting. We also observe how the joint accuracy falls compared to the isolated language variety and gender results. This manifests the difficulty of this joint classification task, which continues being an open problem.

Our final comments are to analyse the difference in difficulty of this shared task depending on the task and language. Identifying gender is clearly more difficult than language variety. Despite the first task has only two possible labels, gender differences are generally more subtle and require more context and topic understanding. In contrast, the language variety peculiarities are both differentiable using lexical and semantic aspects of text. These lexical and semantic aspects are also the cause of the differences in function of the language. English and Arabic varieties are more similar at lexical level than Portuguese or Spanish ones. However, the low number of Portuguese varieties employed in this work affects too. Finally, considering the high number of Spanish varieties and its high results, we also consider that some languages have tweets with topics more indicative of the variety, e.g. topics about politics or events.

5 Conclusions

In this work we presented the system designed by Symanto Research for the PAN 2017 author profiling shared task. The pipeline we present in this paper is easily replicable and yields a good performance while promising to be robust and flexible in the presence of noisy data.

We described an approach based on subword character n-gram embeddings and deep averaging networks. We explained the rationale behind using these components in author profiling. We compared our approach with several well-known baseline models.

Task	Model	Arabic	English	Portuguese	Spanish	Average
Language variety	Random	25.0	16.7	50.0	14.3	26.5
	BOW	71.2	59.4	88.7	75.1	73.6
	Skip-gram emb.	73.0	62.4	98.6	80.6	78.7
	Subword emb.	70.7	68.3	98.5	79.6	79.3
	DAN	**80.6**	**76.5**	**98.9**	**91.0**	**86.8**
Gender	Random	50.0	50.0	50.0	50.0	50.0
	BOW	66.4	66.7	71.0	63.4	66.9
	Skip-gram emb.	71.2	78.4	76.5	73.3	74.8
	Subword emb.	73.7	78.8	72.6	74.5	74.9
	DAN	**74.5**	**80.8**	**78.8**	**75.5**	**77.4**

Table 2: Classification accuracy (in %) using 10-fold cross-validation with the training partition.

Task	Model	Arabic	English	Portuguese	Spanish	Average
Language variety	Random	25.0	16.7	50.0	14.3	26.5
	DAN	**76.6**	**75.9**	**97.9**	**90.0**	**85.1**
Gender	Random	50.0	50.0	50.0	50.0	50.0
	DAN	**73.0**	**79.6**	**76.9**	**77.2**	**76.7**
Joint	Random	12.5	8.4	25.0	7.2	13.3
	DAN	**56.9**	**60.5**	**75.3**	**70.2**	**65.7**

Table 3: Test classification accuracy (in %).

Experimental results in the tasks of native language and gender identification show the superiority of our approach and demonstrate that it is a competitive alternative.

Future work will investigate further how to employ semantic representations and deep learning techniques in the task of author profiling.

References

1. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Communications of the ACM 52(2), 119–123 (2009)
2. Bayot, R., Gonçalves, T.: Author Profiling using SVMs and Word Embedding Averages—Notebook for PAN at CLEF 2016. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal. CEUR-WS.org (Sep 2016)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
4. 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. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal. CEUR-WS.org (Sep 2016)
5. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research 12(Aug), 2493–2537 (2011)
6. Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B.: Tat: an author profiling tool with application to arabic emails. In: Proceedings of the Australasian Language Technology Workshop. pp. 21–30 (2007)
7. Franco-Salvador, M., Rangel, F., Rosso, P., Taulé, M., Martí, M.A.: Language variety identification using distributed representations of words and documents. In: Proceeding of the 6th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction (CLEF 2015). vol. LNCS(9283). Springer-Verlag (2015)
8. Franco-Salvador, M., Rosso, P., Rangel, F.: Distributed representations of words and documents for discriminating similar languages. In: Proceeding of the RANLP Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial). Hissar, Bulgaria (2015)
9. Hahnloser, R.H., Sarpeshkar, R., Mahowald, M.A., Douglas, R.J., Seung, H.S.: Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature 405(6789), 947–951 (2000)
10. Iyyer, M., Manjunatha, V., Boyd-Graber, J., Daumé III, H.: Deep unordered composition rivals syntactic methods for text classification. In: Association for Computational Linguistics (2015)
11. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Koppel, M., Schler, J.: Exploiting stylistic idiosyncrasies for authorship attribution. In: Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis. vol. 69, p. 72 (2003)
13. Koppel, M., Schler, J., Zigdon, K.: Automatically determining an anonymous author's native language. In: Intelligence and Security Informatics, pp. 209–217. Springer (2005)
14. López-Monroy, A.P., y Gómez, M.M., Escalante, H.J., Villaseñor-Pineda, L., Stamatatos, E.: Discriminative subprofile-specific representations for author profiling in social media. Knowledge-Based Systems 89, 134 – 147 (2015)
15. Maier, W., Gómez-Rodríguez, C.: Language variety identification in spanish tweets. In: Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants. pp. 25–35. Doha, Qatar (October 2014)
16. Malmasi, S., Dras, M.: Language identification using classifier ensembles. In: Proceeding of the RANLP Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial). Hissar, Bulgaria (2015)
17. 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 3rd Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (VarDial). Osaka, Japan (2016)
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the Annual Neural Information Processing (NIPS'13) Conference - Advances in Neural Information Processing Systems 26. pp. 3111–3119 (2013)
19. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France. CEUR-WS.org (Sep 2015)

20. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation for language variety identification. In: Proceedings of the 17th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2016). Springer-Verlag (2016)
21. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd Author Profiling Task at PAN 2014. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK. CEUR-WS.org (Sep 2014)
22. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain (Sep 2013)
23. Rangel, F., Rosso, P., Potthast, M., Stein, B.: In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs
24. 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: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2016)
25. Sadat, F., Kazemi, F., Farzindar, A.: Automatic identification of arabic language varieties and dialects in social media. In: In Proceeding of the 1st. International Workshop on Social Media Retrieval and Analysis SoMeRa (2014)
26. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. vol. 6, pp. 199–205 (2006)
27. Tetreault, J., Blanchard, D., Cahill, A.: A report on the first native language identification shared task. In: Proceedings of the eighth workshop on innovative use of NLP for building educational applications. pp. 48–57. Citeseer (2013)
28. Zampieri, M., Gebre, B.G.: Automatic identification of language varieties: The case of Portuguese. In: KONVENS2012-The 11th Conference on Natural Language Processing. pp. 233–237. Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI) (2012)
29. Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J.: A report on the DSL Shared Task 2014. In: Proceedings of the COLING First Joint Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial). pp. 58–67. Dublin, Ireland (August 2014)
30. Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J., Nakov, P.: Overview of the DSL Shared Task 2015. In: Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial). Hissar, Bulgaria (2015)