Author Profiling with Word+Character Neural Attention Network

Notebook for PAN at CLEF 2017

Yasuhide Miura, Tomoki Taniguchi, Motoki Taniguchi, and Tomoko Ohkuma

Fuji Xerox Co., Ltd.
{yasuhide.miura, taniguchi.tomoki, motoki.taniguchi, ohkuma.tomoko}@fujixerox.co.jp

Abstract This paper describes neural network models that we prepared for the author profiling task of PAN@CLEF 2017. In previous PAN series, statistical models using a machine learning method with a variety of features have shown superior performances in author profiling tasks. We decided to tackle the author profiling task using neural networks. Neural networks have recently shown promising results in NLP tasks. Our models integrate word information and character information with multiple neural network layers. The proposed models have marked joint accuracies of 64–86% in the gender identification and the language variety identification of four languages.

1 Introduction

Researches to automatically extract author profile traits from social media have been done to empower activities such as advertisement, forensic, marketing, personalization, and security. PAN tasks have focused on traits like gender, age, and personality type in the past series. This year's author profiling task was to identify a gender and a language variety of a Twitter user [15]. In the gender identification, a task participant is required to determine whether a user is male or female from tweets. Similar gender identifications have been done in past PAN series with different native languages and domains. In the language variation identification, a task participant has to decide a language variety within a given native language from tweets. The study of language varieties has been done in VarDial shared tasks[17] targeting journalistic texts, but is new in PAN series targeting Twitter texts.

Statistical models using a machine learning method like support vector machine have shown effectiveness to identify profile traits in past PAN series. Various features were introduced to these models including word n-grams[6,12,3], character n-grams[6,12,3], part-of-speech tags[6,3], styles[6,12,3], and second order attributes[6]. We decided to tackle the identifications of gender and language variety using neural networks. Neural networks have shown effectiveness to capture complex representations combing simpler representations[9]. We aim to obtain complex representations that were expressed as independent features in the past studies using neural networks. Neural networks such as multilayer perceptron and restricted Boltzmann machine have been used in PAN 2016[16] to obtain word embeddings[2] and as a classifier. Our models combine word information and character information with complex neural networks

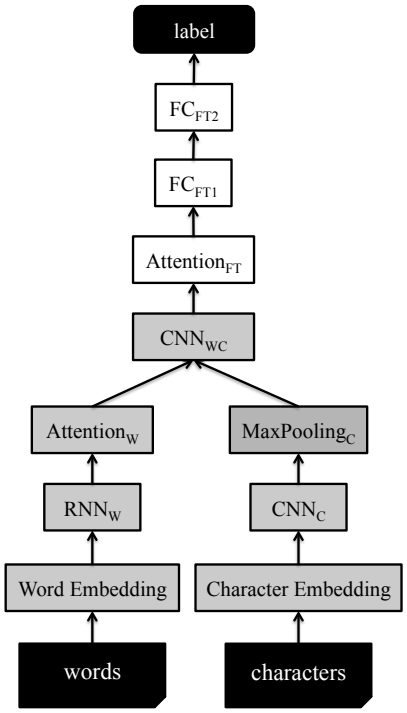

Figure 1. The architecture of model NN-FT. The shaded layers are tweet-level processes.

consisting of a recurrent neural network layer, a convolutional neural network layer, and an attention mechanism[1] layer to classify a profile trait.

In the following section of this paper, we first describe our neural network models in Section 2. Data used in the models are explained in Section 3 following Section 4 with the details of an experiment to confirm the performances of the models. Finally, Section 5 concludes the paper with some future directions.

2 Models

We propose two models that consist of multiple layers to classify a profile trait with neural networks. The architectures of the two models share most of their layers but differ in the fusion strategies of word information and character information. The first model NeuralNet-FusionTweet (NN-FT) combines the two kinds of information with a tweet-level fusion. The second model NeuralNet-FusionUser (NN-FU) performs a fusion process in user-level.

2.1 Model NN-FT

Figure 1 shows the architecture of NN-FT. For each user, the model accepts the words and the characters of user tweets. Note that the words and the characters are just dif-

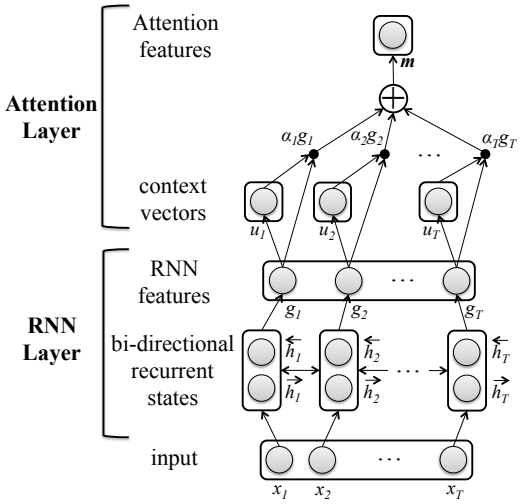

Figure 2. Overview of word processes with $\mathrm{RNN_W}$ and $\mathrm{Attention_W}$.

ferent representations of same tweet texts. The words and the characters are embedded with embedding layers and are processed with a recurrent neural network (RNN) layer, convolutional neural network (CNN) layers, attention mechanism[1] layers, a max-pooling layer, and fully-connected (FC) layers. As an implementation of RNN, we used Gated Recurrent Unit (GRU)[7] with a bi-directional setting.

word processes Figure 2 illustrates the overview of word processes by $\mathrm{RNN_W}$ and $\mathrm{Attention_W}$. The input words are embedded to k_w dimension word embeddings with embedding matrix $\boldsymbol{E}_w$ to obtain $\boldsymbol{x}$ with $\boldsymbol{x}_t \in \mathbb{R}^{k_w}$. $\boldsymbol{x}$ are then processed in $\mathrm{RNN_W}$ with the following transition functions:

$$\boldsymbol{z}_t = \sigma\left(\boldsymbol{W}_z \boldsymbol{x}_t + \boldsymbol{U}_z \boldsymbol{h}_{t-1} + \boldsymbol{b}_z\right) \tag{1}$$

$$\boldsymbol{r}_t = \sigma\left(\boldsymbol{W}_r \boldsymbol{x}_t + \boldsymbol{U}_r \boldsymbol{h}_{t-1} + \boldsymbol{b}_r\right) \tag{2}$$

$$\tilde{\boldsymbol{h}}_t = \tanh\left(\boldsymbol{W}_h \boldsymbol{x}_t + \boldsymbol{U}_h \left(\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}\right) + \boldsymbol{b}_h\right) \tag{3}$$

$$\boldsymbol{h}_t = \left(\boldsymbol{1} - \boldsymbol{z}_t\right) \odot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_t \tag{4}$$

where $\boldsymbol{z}_t$ is an update gate, $\boldsymbol{r}_t$ is a reset gate, $\tilde{\boldsymbol{h}}_t$ is a candidate state, $\boldsymbol{h}_t$ is a state, $\boldsymbol{W}_z, \boldsymbol{W}_r, \boldsymbol{W}_h, \boldsymbol{U}_z, \boldsymbol{U}_r, \boldsymbol{U}_h$ are weight matrices, $\boldsymbol{b}_z, \boldsymbol{b}_r, \boldsymbol{b}_h$ are bias vectors, σ is a logistic sigmoid function, and $\odot$ is an element-wise multiplication operator. The bi-directional GRU outputs $\overrightarrow{\boldsymbol{h}}$ and $\overleftarrow{\boldsymbol{h}}$ are concatenated to form $\boldsymbol{g}$ $where$ $\boldsymbol{g}_t = \overrightarrow{\boldsymbol{h}_t} \| \overleftarrow{\boldsymbol{h}_t}$ and are passed to $\mathrm{Attention_w}$.

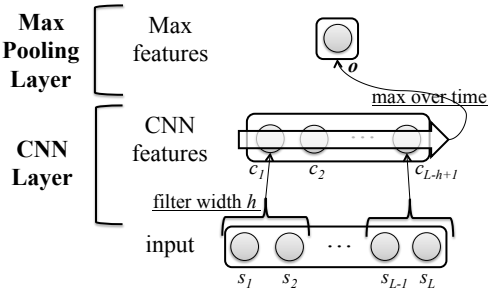

Figure 3. Overview of character processes with $\mathrm{CNN_C}$ and $\mathrm{MaxPooling_C}$.

$\mathrm{Attention_W}$ computes a tweet representation $\boldsymbol{m}$ as a weighted sum of $\boldsymbol{g}_t$ with weight α_t:

$$\boldsymbol{m} = \sum_t \alpha_t \boldsymbol{g}_t \tag{5}$$

$$\alpha_t = \frac{\exp\left(\boldsymbol{v}_\alpha^T \boldsymbol{u}_t\right)}{\sum_t \exp\left(\boldsymbol{v}_\alpha^T \boldsymbol{u}_t\right)} \tag{6}$$

$$\boldsymbol{u}_t = \tanh\left(\boldsymbol{W}_\alpha \boldsymbol{g}_t + \boldsymbol{b}_\alpha\right) \tag{7}$$

where $\boldsymbol{v}_\alpha$ is a weight vector, $\boldsymbol{W}_\alpha$ is a weight matrix, and $\boldsymbol{b}_\alpha$ a bias vector. $\boldsymbol{u}_t$ is an attention context vector calculated from $\boldsymbol{g}_t$ with a single FC layer (Eq. 7). $\boldsymbol{u}_t$ is normalized with softmax to obtain α_t as a probability (Eq. 6).

character processes Figure 3 illustrates the overview of character processes by $\mathrm{CNN_C}$ and $\mathrm{MaxPooling_C}$. The input characters are embedded to k_c dimension character embeddings with character embedding matrix $\boldsymbol{E}_c$ to obtain $\boldsymbol{s}$ with $\boldsymbol{s}_i \in \mathbb{R}^{k_c}$. $\boldsymbol{s}$ is then passed to $\mathrm{CNN_C}$ to obtain $\boldsymbol{c}$ with:

$$c_i = f\left(\boldsymbol{W}_c \boldsymbol{s}_{i:i+h-1} + \boldsymbol{b}_c\right) \tag{8}$$

where $f(\cdot)$ is a non-linear function, $\boldsymbol{W}_c$ is a weight matrix, h a convolution window size, and $\boldsymbol{b}_c$ a bias vector. We used rectified linear unit for $f(\cdot)$. $\boldsymbol{c}$ is further processed with max-over time process[8] in $\mathrm{MaxPooling_C}$ to obtain a tweet representation $\boldsymbol{o}$.

word+character processes Two tweet representations $\boldsymbol{m}$ and $\boldsymbol{o}$ are concatenated to further apply word+character processes. The concatenated tweet representation is processed by $\mathrm{CNN_{WC}}$ like in $\mathrm{CNN_C}$ with window size $h = 1$ to get a word and character combined representation. The combined tweet representation is then passed to $\mathrm{Attention_{FT}}$ to obtain a user representation from tweet representations. Finally, the user representation is passed to $\mathrm{FC_{FT1}}$ and $\mathrm{FC_{FT2}}$, respectively.

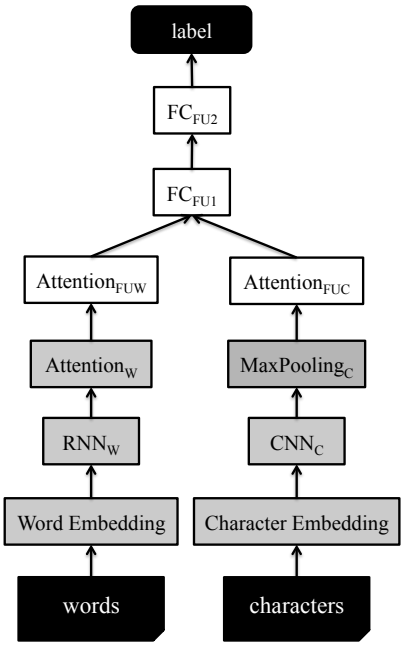

Figure 4. The architecture of model NN-FU. The shaded layers are tweet-level processes.

2.2 Model NN-FU

Figure 4 shows the architecture of NN-FU. Many layers in NN-FU exist in NN-FT. Layers that are not apparent in NN-FT are $\mathrm{Attention}_{\mathrm{FUW}}$, $\mathrm{Attention}_{\mathrm{FUC}}$, $\mathrm{FC}_{\mathrm{FU1}}$, and $\mathrm{FC}_{\mathrm{FU2}}$. $\mathrm{Attention}_{\mathrm{FUW}}$ merges tweet representations obtained from word information. Similarly, $\mathrm{Attention}_{\mathrm{FUC}}$ merges tweet representations obtained from character information. The outputs of these attention processes are concatenated and is further processed with $\mathrm{FC}_{\mathrm{FU1}}$ and $\mathrm{FC}_{\mathrm{FU2}}$.

The attention processes in NN-FU are different from the attention processes in NN-FT, where word information and character information are concatenated prior to $\mathrm{Attention}_{\mathrm{FT}}$. In NN-FU, word information and character information are concatenated after the attention processes with user-level representations. The other non-apparent layers $\mathrm{FC}_{\mathrm{FU1}}$ and $\mathrm{FC}_{\mathrm{FU2}}$ perform similarly as $\mathrm{FC}_{\mathrm{FT1}}$ and $\mathrm{FC}_{\mathrm{FT2}}$ in NN-FT to process a word+character user representation.

3 Data

The weights in the proposed models require some data to be trained. We used two datasets to train the proposed models with two different objectives.

Languages		English, Spanish, Portuguese, Arabic
Gender Labels		male, female
Language Variety Labels	English	Australia, Canada, Great Britain, Ireland, New Zealand, United States
	Spanish	Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela
	Portuguese	Brazil, Portugal
	Arabic	Egypt, Gulf, Levantine, Maghrebi

Table 1. The languages, the gender labels, and the language variety labels of PAN@CLEF 2017 Author Profiling Training Corpus.

Language	#tweet
English	10.72M
Spanish	3.17M
Portugese	2.75M
Arabic	2.46M

Table 2. The number of tweets collected for each language with Twitter Streaming APIs. M in the table represents the million unit.

3.1 PAN@CLEF 2017 Author Profiling Training Corpus

The first dataset we used to train the proposed models is the official PAN@CLEF 2017 Author Profiling Training Corpus. The dataset consists of $11,400$ Twitter users in four languages with the gold labels of gender and language variety. The languages, gender labels, and language variety included in this dataset is summarized in Table 1 This dataset is used to train the models to minimize an empirical loss between predictions and gold labels.

We divided this dataset into $train_8$, dev_1, and $test_1$ with a stratified sampling by ratio of 8:1:1. These subsets were made so that we can empirically decided some parameters of the models. We will describe the detail of parameter selection in Section 4.2.

3.2 Streaming Tweets

The second dataset we used to train the proposed models is tweets collected by Twitter Streaming APIs[1]. We collected these tweets to pre-train the word embedding matrix E_w of the models. Neural network models are known to perform better when word embeddings are pre-trained by a large-scale dataset[8]. The following steps describe the detail of the collection process:

1. Tweets with lang metadata of en, es, pt, and ar were collected via Twitter Streaming APIs during the period of March–May 2017.

[1] https://dev.twitter.com/streaming/overview

Parameter	Size
word embedding dimension	100
character embedding dimension	25
$\mathrm{RNN_W}$ units	100
$\mathrm{CNN_C}$ units	50
$\mathrm{CNN_{WC}}$ units	300
$\mathrm{CNN_C}$ filter sizes	$3, 6$
$\mathrm{CNN_{WC}}$ filter size	1
$\mathrm{Attention_W}$ units	200
$\mathrm{Attention_{FT}}$ units	300
$\mathrm{Attention_{FUW}}$ units	200
$\mathrm{Attention_{FUC}}$ units	100
$\mathrm{FC_{FT1}}$ units	150
$\mathrm{FC_{FU1}}$ units	150
$\mathrm{FC_{FT2}}$ units	#label
$\mathrm{FC_{FU2}}$ units	#label

Table 3. The sizes of parameters in the proposed models.

2. Retweets are removed from the collected tweets.
3. Tweets posted by bots[2] are deleted from the collected tweets.

Table 2 shows the number of resulting tweets. We will describe the detail of word embedding pre-training in Section 4.1.

4 Experiment

4.1 Model Configurations

Text Processor We applied a unicode normalization, a Twitter user name normalization, and a URL normalization for text pre-processing. Pre-processed texts were tokenized with the two kinds of tokenizers. Twokenizer[13] is used for English and NLTK[4] WordPunctTokenizer is used for other languages. Words are converted to lower case forms to ignore capitalization. Note that the lower case conversion is not performed for character inputs.

Initialization of Embeddings We pre-trained word embeddings using streaming tweets of Section 3.2 by fastText[5] with the skip-gram algorithm. The pre-training parameters are dimension=100, learning rate=0.025, window size=5, negative sample size=5, and epoch=5. For character embeddings, we randomly initialized them with a uniform distribution.

Convolution Filter Sizes, Layer Unit Sizes, and Word Embedding Sizes Table 3 summarizes the sizes of various parameters included in the proposed models. In $\mathrm{CNN_C}$, two values are listed since we used the multiple filters approach[10].

[2] We assembled a Twitter client list consisting of 80 clients that are used for manual postings.

Language	NN-FT		NN-FU	
	Accuracy	α	Accuracy	α
English	80.00	1e-4	81.94	5e-4
Spanish	79.52	5e-5	77.62	5e-6
Portuguese	84.17	5e-5	90.83+	5e-7, 1e-7
Arabic	76.25	1e-3	79.17	5e-4

Table 4. Gender identification results of the proposed models on $test_1$. $^+$ values are averaged values.

Optimization Strategy We used cross-entropy loss as an objective function of the models. l_2 regularization was applied to the RNN layers, the attention context vectors, the CNN layers, and the FC layers of the models to avoid overfitting. The objective function was minimized through stochastic gradient descent over shuffled mini-batches with Adam[11]. For the initial learning rate of Adam, we set it to $1e^{-3}$.

Parameter Selection The models have regularization parameter α which is sensitive to a dataset. We selected optimal values for α:

$$\alpha \in \left\{1e^{-3}, 5e^{-4}, 1e^{-4}, 5e^{-5}, 1e^{-5}, 5e^{-6}, 1e^{-6}, 5e^{-7}, 1e^{-7}\right\}$$

in terms of accuracy with a grid search using dev_1 described in Section 3.1.

4.2 In-house Experiment

We evaluated the proposed models using $train_8$, dev_1, and $test_1$. All models are trained using a single NVIDIA Titan X gpu. Table 4 presents the results of gender identifications. In the gender identifications, NN-FU performed better than NN-FT with one exception in Spanish. Table 5 shows the results of language variety identifications. The language variety identifications showed different characteristics where NN-FT performing better in all languages compared to NN-FU.

4.3 Submission Run

We chose the best performing models and αs in the in-house experiment as models and parameters for a submission run. In the cases of multiple best performing αs, we chose αs that showed the best performances in $test_1$. The submission run was done in a TIRA virtual machine [14] with cpus. Table 6 summarizes the performances of the models in the submission run. The models showed a similar trend as in the in-house experiment. They ranked 3rd in gender ranking, 6th in language variety ranking, and 4th in the global ranking.

Language	NN-FT		NN-FU	
	Accuracy	α	Accuracy	α
English	85.83^{+}	$5e^{-7}$, $1e^{-7}$	85.56	$1e^{-6}$
Spanish	93.65^{+}	$1e^{-4}$, $1e^{-5}$, $1e^{-7}$	93.10^{+}	$5e^{-4}$, $1e^{-7}$
Portuguese	99.89^{+}	$1e^{-3}$, $5e^{-4}$, $1e^{-4}$, $5e^{-5}$, $1e^{-5}$, $5e^{-6}$, $1e^{-6}$, $5e^{-7}$, $1e^{-7}$	99.47^{+}	$1e^{-3}$, $5e^{-4}$, $1e^{-4}$, $5e^{-5}$, $1e^{-6}$, $5e^{-7}$, $1e^{-7}$
Arabic	78.33^{+}	$1e^{-6}$, $1e^{-7}$	77.08	$1e^{-3}$

Table 5. Language variety identification results of the proposed models on $test_1$. $^{+}$ values are averaged values.

Language	Trait	Model	Accuracy	Joint Accuracy
English	gender	NN-FU	80.46	69.92
	language variety	NN-FT	87.17	
Spanish	gender	NN-FT	81.18	75.18
	language variety	NN-FT	92.71	
Portuguese	gender	NN-FU	87.00	85.75
	language variety	NN-FT	98.13	
Arabic	gender	NN-FU	76.44	64.19
	language variety	NN-FT	81.25	

Table 6. The performances of the proposed models in the submission run.

5 Conclusion

As described in this paper, we proposed two models, NN-FT and NN-FU, for author profiling. The two models differ in the fusion strategies of word information and character information. The models marked joint accuracies of 64–86% in the gender identification and the language variety identification of four languages. They performed better in gender identification compared to language variety identification. The average accuracies from the top systems were -1.26% for gender and -2.05% for language variety. This result is not so surprising since neural network models had shown difficulties adapting to language variety identification in past VarDial shared tasks [17].

As future works of this study, we plan to analyze differences of internal states in NN-FT and NN-FU. The best performing models were different among profile traits and languages in the in-house experiment. We will like to unveil the causes of this differences to further improve our models.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. Computing Research Repository abs/1409.0473 (2014), http://arxiv.org/abs/1409.0473
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 (2016)
3. Bilan, I., Zhekova, D.: CAPS: A Cross-genre Author Profiling System—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 (2016)
4. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc. (2009)
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
6. 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 (2016)
7. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734 (2014)
8. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research 12, 2493–2537 (2011)
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
10. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751 (2014)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. Computing Research Repository abs/1412.6980 (2014), http://arxiv.org/abs/1412.6980
12. Modaresi, P., Liebeck, M., Conrad, S.: Exploring the Effects of Cross-Genre Machine Learning for Author Profiling in PAN 2016—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 (2016)
13. Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., Smith, N.A.: Improved part-of-speech tagging for online conversational text with word clusters. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT). pp. 380–390 (2013)
14. 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). pp. 268–299 (2014)
15. 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 (2017)

16. Rangel Pardo, 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 (2016)
17. Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., Aepli, N.: Findings of the vardial evaluation campaign 2017. In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). pp. 1–15 (2017)