Multilingual Gender Classification with Multi-view Deep Learning

Notebook for PAN at CLEF 2018

Matej Martinc[1,2], Blaž Škrlj[1,2], and Senja Pollak[1,3]

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[3] USHER Institute, University of Edinburgh, United Kingdom
{matej.martinc,blaz.skrlj,senja.pollak}@ijs.si

Abstract We present the results of a gender identification performed on the data set of tweets and images prepared for the PAN 2018 Author profiling shared task. In this work we propose a hybrid neural network architecture for gender classification, capable of leveraging heterogeneous textual and image information sources. The proposed approach is based on state-of-the-art deep architectures for natural language processing, combined with a pretrained image classification architecture via a custom output combination scheme. Text classification model combines character level, word level and document level information in order to produce stable and accurate predictions on three different languages, achieving the highest accuracy of 79% on the English test set. Image classification architecture relies on the hypothesis that the authors have a gender bias when it comes to publishing images of people and has a structure of a two-phased pipeline containing two models, one for face detection and the other for face gender classification. Classifying author's gender from posted images proved to be harder than from text, with our image classification model achieving the best accuracy of only 58.26% on the English test set. The results on the official PAN test set also confirmed slight synergy effects between the two models when combined. The proposed approach was 8th in the global ranking of PAN 2018 Author profiling shared task.

1 Introduction

The heterogeneous image and text data from social media has become a popular resource for studies in data mining, especially due to its accessibility, size and a near real-time publishing. The trend of publishing content describing personal experiences, stands and emotions and the sheer size of the data available has allowed the development of statistical models, capable of determining the users' characteristics related to demographics, psychological profile and mental health. The field that deals with discovering users' attributes from this content automatically is known as author profiling (AP) and includes tasks, such as the prediction of author's gender [16], age [18], personality type [25] or language variety [27].

In order to encourage further development and sharing of methods and results from the field of AP, a series of scientific events and shared tasks on digital text forensic called

PAN (Uncovering Plagiarism, Authorship, and Social Software Misuse) [24] have been organized. The first PAN event took place in 2011, while the first AP shared task was organized in 2013 [15]. Traditionally, AP approaches in PAN shared task used only textual data, such as tweets and similar personal publications to develop classifiers. State-of-the-art approaches on this type of data mostly relied on traditional classifiers and required extensive feature engineering [17]. However, with the latest improvements in image recognition methodology, the PAN 2018 Author profiling challenge [16] offered for the first time an opportunity to leverage also image material. This paradigm shift prompted us to use a novel neural network architecture capable of leveraging different sources of information to maximize performance and yield robust results. Therefore, in this work we investigate how current state-of-the-art text-based deep learning architectures can be combined with a set of recently introduced image preprocessing and classification techniques.

2 Related Work

The earliest attempts at AP that covered gender identification started with [7], who used parts of the BNC, but continued on other corpora, such as the ACL corpus of scientific papers [26] and more recently the social media corpora. The best gender profiling approaches within the last year's PAN shared task on tweets [17] achieved the accuracy of 0.8233 for English and 0.8321 for Spanish. The approach was proposed by Basile et al. [2] who had also the overall best results with combinations of character and tf-idf word n-grams trained with an SVM. For Arabic, the highest score of 0.8031 was a result of our system [11], which used a combination of word and different types of character n-grams [20], as well as POS n-grams, sentiment from emojis and character flooding as features in Logistic regression classifier.

Some neural networks approaches were also proposed and for Portuguese the highest results (87%) were achieved by [13], using an architecture consisting of a recurrent neural network layer, a convolutional neural network (CNN) layer, and an attention mechanism [1] layer capable of integrating character and word information. Two other deep learning approaches ([22] and [21]) also laveraged CNN architecture but overall achieved worse results.

For gender identification, also online workflows have been proposed [10] in the ClowdFlows environment [8], including workflows for fast experimentation when training new gender models[4], use of pretrained gender classifiers for five languages (based on [11]), which can be used for example for linguistic analysis[5] or for evaluating models on new datasets (e.g., in the cross-genre evaluation setting. However, this workflows currently do not support deep learning architectures.

When it comes to predicting gender from images, all state-of-the-art approaches deploy neural architecture. One of the more recent approaches is the one proposed by [9], which showed that a simple CNN architecture can be successfully employed for gender and age classification even when the amount of learning data is limited. Another

[4] `http://clowdflows.org/workflow/10620/`

[5] `http://clowdflows.org/workflow/10980/`

successful example of employing neural architecture in the field of AP is an age classification model proposed by [19], which is described in more detail in Sections 3 and 4.2, since its adaptation is also used in this paper.

The dataset of this year's task is multimodal, therefore we also researched some multi-view learning approaches. Multi-view learning concerns with leveraging different data sources to learn a more complete representation of the modeled system. It has been an active area of research for more than 15 years. Modern multi-view learning approaches face two main issues: scalability and the method for view combination. Recent multi-view improvements in the area of deep learning include for example: recommender systems, where they have shown different sources of information across multiple domains can be used to produce better predictive models [5]. This work builds on the current state-of-the-art approaches for multi-view deep learning by combining predictions learned using text, as well as images, using a novel combination scheme based on preliminary imperical tests.

3 Data Set Description and Preprocessing

Official PAN 2018 AP train set consists of tweets in three different languages grouped by tweet authors, who are labeled by gender (Table 1). Data set is balanced which means that half of the authors in every language are male and half are female. This train set was used in our experiments for parameter tuning and training of the classification models.

Table 1. PAN 2018 training set structure

Language	Authors	Tweets	Images
English	3,000	300,000	30,000
Spanish	3,000	300,000	30,000
Arabic	1,500	150,000	15,000

Text preprocessing was light, for English and Spanish consisting only of replacing all hashtags, mentions and URLs with specific placeholders #HASHTAG, @MENTION, HTTPURL, respectively. For Arabic, an additional step of reversing tweets was performed since they are written from right-to-left. Finally, all tweets belonging to the same author were concatenated and used as one document in further processing.

Preliminary experiments were used in order to decide on the most effective image preprocessing technique. Images from all the English authors were labeled with gender labels corresponding to their authors and split into a train set (80% of images) and validation set (20% of images). A baseline deep architecture model was used for classifying individual images in the validation set. We experimented with different CNN configurations of up to the depth of 10 layers, where different combinations of activations were used. Such direct approach merely outperformed the 50% baseline classifier. Because of this we decided to only use images containing human faces. The hypothesis was that male users post more images of men and female users post more female images (possibly also due to selfies).

After the images were initially rescaled to 64×64 pixels, a face detection algorithm, originally proposed by [19], was used to extract the images, containing one or more faces. The used DEX (Deep EXpectation) works as follows. First, faces corresponding to the same person are aligned using an explicit alignment algorithm, which considers angles between $\pm + 60^\circ$ and at the $\pm 90^\circ$ angle—to handle rotated images. The algorithm is based on the work done by [12] The authors of the original study demonstrated, that this approach is more robust when compared to baseline landmark detectors. Schematic representation of the face detector is presented in Figure 1.

We leveraged the pre-trained convolutional architecture for face extraction (freely accessible at `https://github.com/yu4u/age-gender-estimation`) and combined it with our baseline deep architecture model for individual image classification. This architecture achieved 62% accuracy when trained on the same train-validation English image data set split as before, confirming our thesis that authors post more images of people of the same gender.

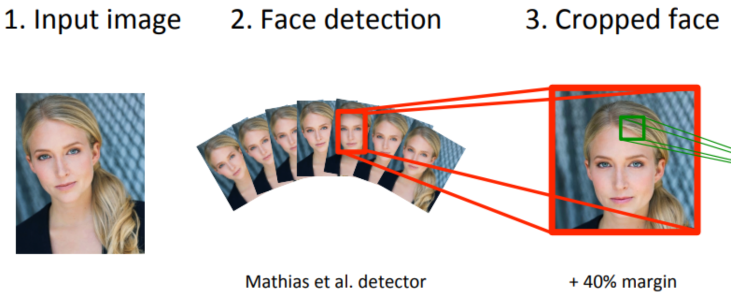

Figure 1. Schematic representation of representative image construction, as used in [19]

4 Classification Model

Our gender classification model consists of two separate neural network models, one for text classification and one for image classification. Two models are combined in the final classification step in order to produce a unified prediction for every author.

4.1 Text Classification Model

The final text classification model (visualized in Figure 2) is a combination of three distinct text classification architectures, capable of leveraging character level, word level and document level information. This architecture proved to be a good substitute for extensive feature engineering usually applied in AP classification tasks and worked reasonably well even when trained on a relatively small train set. Preprocessed text is fed to the network presented in Figure 2 in the form of three distinct inputs:

- *Char sequences*: Every preprocessed document is converted into a numeric char sequence (every char is represented by a distinct integer) of length corresponding to the number of chars in the longest document in the train set (zero value padding is added after the document char sequence and truncating is also performed at the end of the sequence).
- *Word sequences*: Every preprocessed document is tokenized and words which appear in less than 30% or in more than 70% of documents from the train set are removed. The resulting word sequences are converted into integer sequences of length corresponding to the length of the longest sequence (again zero value padding is used but this time padding is added at the beginning of the sequence).
- *TF-IDF matrix*: Preprocessed input data set is converted into a matrix of TF-IDF features with a TfidfVectorizer from ScikitLearn [14]. The matrix is calculated on lowercased word unigrams with a minimum document frequency of 10 and appearing in at most 80% of the documents in the train set. Sublinear term frequency scaling is applied in the term frequency calculation.

The architecture for processing *Word sequences* follows the approach proposed by [6], consisting of a CNN model in addition to pre-trained word vectors (we use pre-trained FastText embeddings [3] of size 300. A distinct feature c_i is produced for every possible window of h words $x_{i:i+h-1}$ in the document according to the convolutional equation:

$$c_i = f(w \cdot x_{i:i+h-1} + b)$$

where *w* is a convolutional filter, *b* is a bias and *f* a non-linear function (a rectified linear unit (*ReLU*) in our case). A max-over-time pooling operation [4] is applied on the resulting set of features generated for every window size in order to get the most important feature (the one with the highest predictive power). All convolutional filters are of size 64 and windows of sizes 3, 4, 5, 10, 30 and 50 are used. The vectors of the most important features are then concatenated and the previously described convolutional equation together with the max-over-time pooling is applied again on the concatenated features. Finally, the resulting output is passed to a fully connected (*dense*) layer.

The architecture for processing *Char sequences* follows a very similar general idea. The main differences are that character embeddings are not pretrained, windows of sizes 2,3 and 4 are used and two additional layers (one convolutional and one max-over-time pooling layer) were added before the fully connected layer.

The processing of the *TF-IDF matrix* is more straight forward. The matrix is first passed to a fully connected layer of size 128. We conduct a dropout operation on the output of the layer, in which 40% of input units are dropped in order to reduce overfitting, and *ReLU* is employed on the remaining units. Finally, the resulting output is again passed to a *dense* layer.

The output of the three resulting *dense* layers (one for every input) are concatenated, dropout is conducted on the concatenation, and *ReLU* is employed on the remaining units. A final step in the text classification model is passing the resulting vectors to a dense layer with a *sigmoid* activation, whose output is the probability distribution over two gender classes.

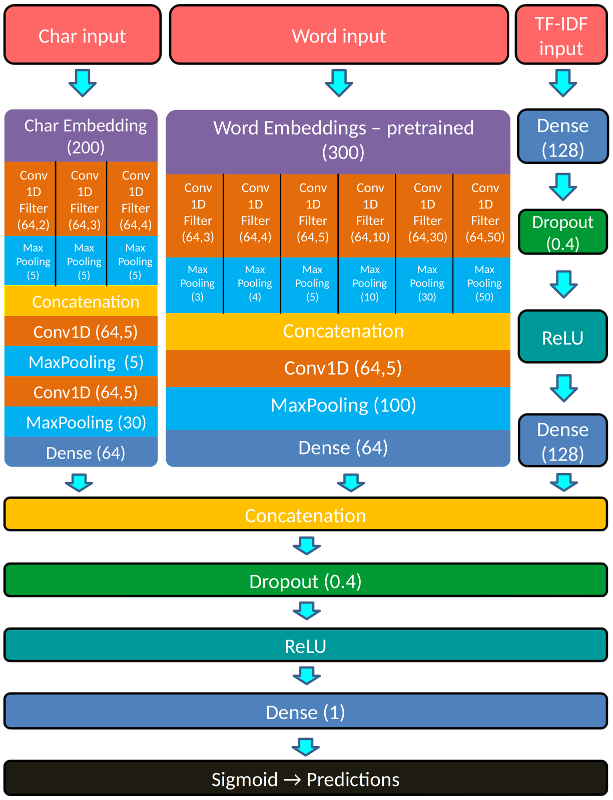

Figure 2. Neural network architecture for text classification.

4.2 Image Classification Model

As was already explained in Section 3, we only used images of faces for gender classification. We hypothesized, that a pre-trained model trained on a big data set, containing fairly reliable gender labels for males and females in the image, would achieve better results than a model trained on our somewhat messy data set of images of people (see Section 3 for a description of how we built this data set), where gender labels corresponded to the gender of the person who published the images and not necessarily to the person on the image itself. This hypothesis was experimentally confirmed and therefore an already trained model published at `https://github.com/yu4u/age-gender-estimation` was used for image classification. This model is an adaptation of the age classification model proposed by [19] and has the same design principles (CNN of VGG-16 architecture [23]). The model was trained on IMDB-WIKI data set (`https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/`) and is capable of assigning gender labels to one or more faces in the input image. For classifying gender of the author from his/her posted images, the following procedure is applied. First, we check if none of the images contain any faces. If that is the case, we automatically assign male gender (since in a bit more than half of the cases the authors without images with faces are male). If the images posted by the author contain faces, we classify them all and count how many of them were classified as male and how many as female. If there are more female faces than male, the author is classified as female, otherwise as male.

4.3 Combining Image and Text Models

Preliminary experiments were conducted in order to combine text and image classification models in a way that would maximize synergy effects. The text classification model proved considerably more accurate than the image classification model, therefore it was decided to only use image classifiers's prediction if the following three conditions are met:

1. There are at least two faces found in images published by the author.
2. All the faces are of the same gender.
3. The *sigmoid* function output of the text classifier falls between 0.3 and 0.7, signalling non-confident prediction.

5 Results and Discussion

First, we tested our model in a 10-fold cross validation setting on the PAN 2018 train set. For ten times, text classification model is trained on nine folds of the data set and combined with an image classification model, which is pretrained on an IMDB-WIKI data set as explained in 4.2. Text, image and combined classification models are then all tested on one tenth of the data. The results are presented in Table 2.

Table 2. Gender classification accuracy results of 10-fold cross-validation

Language	Text model	Image model	Combined model
Arabic	**0.8209**	0.5687	0.8141
English	0.8150	**0.5772**	**0.8180**
Spanish	0.7903	0.5516	0.7880

Results show that a text classification model outperforms image classification model by a large margin for every language. Best results for text classification were achieved for Arabic (around 82%) and image classification model achieved best results on English (around 58%). We can see that combining text and image models does not improve the results of the text classification model for Arabic and Spanish and the improvement of 0.3% on the English language is marginal.

For the use on the PAN 2018 AP official test set, the text classification models for all languages were not trained on all the train data. Because of the inherited randomness

Table 3. Gender classification accuracy results on the PAN 2018 AP official test set

Language	Text model	Image model	Combined model
Arabic	0.7760	0.5600	0.7780
English	**0.7900**	**0.5826**	**0.7926**
Spanish	0.7782	0.5486	0.7786
Average	0.7814	0.5637	0.7831

of neural models, which might lead to non-convergence of the trained models in some cases, we decided to use models that were validated in the cross-validation process. Therefore, for every language, the model trained on nine folds, which achieved the highest accuracy on the tenth fold, was used. The accuracy scores achieved by this text classification models were 0.8442 for Arab, 0.8595 for English and 0.8080 for Spanish. Image classification model used on the PAN 2018 AP official test set was again trained on the IMDB-WIKI data set. The results on the official PAN 2018 Author profiling test set are presented in Table 3. In general, the accuracy achieved on the official test set is lower than the one achieved in the cross validation, possibly due to overfitting of the used models. The highest accuracy was achieved on English test set and the combined model achieved the best results on all three languages, confirming slight synergy effects of combining the text and image models, even though the improvements are marginal. The proposed approach scored 8th in the overall leaderboard[6].

The analysis of the authors that were correctly classified by the image classification model and incorrectly classified by the text classification model shows that these authors on average posted 7.568 images of faces per author while the average number of posted images of faces for all authors is 7.218 in the PAN 2018 train set. There is also a noticeable gender imbalance in the set of these authors since it is approximately seven times more likely that the true gender of these authors is male. This suggests that our image classification approach is more appropriate for classifying male authors.

6 Conclusions and Future Work

We propose a novel deep learning approach to gender classification from heterogeneous data sources, namely text and images. The proposed approach is a combination of state-of-the-art deep architectures for natural language processing and a pretrained image classification architecture. The final model returns three distinct gender predictions for every author, one based on text data, one based on image data and one based on all the data. The results confirm the superiority of the text classification model in terms of accuracy, since the predictions based on text are far more accurate than the ones based on images for all the languages. Text classification model was tested in a 10-fold cross-validation setting and achieved the highest accuracy on the Arabic data set (82.09%). Image classification model achieved the highest accuracy on the English data set (57.72%). Combining both models resulted in a marginal accuracy improvement on the English data set.

[6] `https://pan.webis.de/clef18/pan18-web/author-profiling.html`

For future work, we plan to focus on improving the results of the image classification model by testing different architectures for image object detection and image captioning, which might prove helpful in extracting topical information from images that could be used for gender prediction.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-gram: New groningen author-profiling model. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. (2017)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
4. 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)
5. Elkahky, A.M., Song, Y., He, X.: A multi-view deep learning approach for cross domain user modeling in recommendation systems. In: Proceedings of the 24th International Conference on World Wide Web. pp. 278–288. International World Wide Web Conferences Steering Committee (2015)
6. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
7. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
8. Kranjc, J., Podpečan, V., Lavrač, N.: ClowdFlows: A cloud based scientific workflow platform. In: Flach, P.A., Bie, T.D., Cristianini, N. (eds.) Proc. of ECML/PKDD (2). LNCS, vol. 7524, pp. 816–819. Springer (2012)
9. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 34–42 (2015)
10. Martinc, M., Pollak, S.: Reusable workflows for gender prediction. In: chair), N.C.C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Paris, France (may 2018)
11. Martinc, M., Škrjanec, I., Zupan, K., Pollak, S.: Pan 2017: Author profiling-gender and language variety prediction. CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers (2017)
12. Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L.: Face detection without bells and whistles. In: European Conference on Computer Vision. pp. 720–735. Springer (2014)
13. Miura, Y., Taniguchi, T., Taniguchi, M., Ohkuma, T.: Author profiling with word+ character neural attention network. Cappellato et al.[13]
14. 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)
15. 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)

16. Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B.: Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018)
17. 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)
18. 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)
19. Rothe, R., Timofte, R., Van Gool, L.: Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision 126(2-4), 144–157 (2018)
20. 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
21. Schaetti, N.: Unine at clef 2017: Tf-idf and deep-learning for author profiling. Cappellato et al.[13] (2017)
22. Sierra, S., Montes-y Gómez, M., Solorio, T., González, F.A.: Convolutional neural networks for author profiling. Working Notes Papers of the CLEF (2017)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
24. Stamatatos, E., Rangel, F., Tschuggnall, M., Kestemont, M., Rosso, P., Stein, B., Potthast, M.: Overview of PAN-2018: Author Identification, Author Profiling, and Author Obfuscation. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J., Soulier, L., Sanjuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 9th International Conference of the CLEF Initiative (CLEF 18). Springer, Berlin Heidelberg New York (Sep 2018)
25. Verhoeven, B., Daelemans, W., Plank, B.: Twisty: A multilingual twitter stylometry corpus for gender and personality profiling. In: LREC (2016)
26. 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)
27. Zampieri, M., Malmasi, S., Ljubesic, 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 2017, Valencia, Spain, April 3, 2017. pp. 1–15 (2017)