Convolutional Neural Networks for Author Profiling

Notebook for PAN at CLEF 2017

Sebastian Sierra[1], Manuel Montes-y-Gómez[2],
Thamar Solorio[3], and Fabio A. González[1]

[1] Computing Systems and Industrial Engineering Dept., Universidad Nacional de Colombia
Bogotá, Colombia
{ssierral, fagonzalezo}@unal.edu.co
[2] Instituto Nacional de Astrofísica, Óptica y Electrónica
Puebla, Mexico
mmontesg@ccc.inoep.mx
[3] Dept. of Computer Science, University of Houston
Houston, TX, 77004
solorio@cs.uh.edu

Abstract Social media data allows researchers to establish relationships between everyday language and people's sociodemographic variables, such as gender, age, language variety or personality. These variables configure social groups, where author profiling attempts to exploit the idea that they share a common language. This work describes our proposed method for the PAN 2017 Author Profiling shared task. We trained separate models for gender and language variety using a Convolutional Neural Network (CNN). We explored parameters such as the size of the input of the network, the size of the convolutional kernels, the number of kernels and the type of input. We found experimentally that sequences of words performed better than sequences of characters as input for the CNN. We obtained $0.66, 0.73, 0.81$ and 0.57 of accuracy in the test partition for English, Spanish, Portuguese and Arabic respectively.

1 Introduction

Author profiling consists of determining a social group of an unknown author [1]. Chambers et al. support this idea with a sociolinguistics observation, where a social group shares a way of speaking and writing, a dialect [3]. Several profile dimensions for characterizing a social group have been considered since then, such as age, gender, native language and personality. The relevance of this task has been recognized for its applications that include forensics, marketing and security concerns.

Last year's PAN Author Profiling (PAN AP) shared task consisted of a cross-genre age and gender prediction task [14]. 22 teams participated using documents in English, Spanish and Dutch. The teams approached this task using two kinds of features, style-based features and content-based features. Style-based features included n-gram frequencies, punctuations, readability. Whereas content-based features comprise bag of words, word n-grams, term vectors, named entities, among others. Previous successful approaches have used style and content features. Argamon et al. showed experimentally that content features performed better for language, age and gender profiling [1].

However these features can be very sparse. López-Monroy et al. approached the profiling task using a low-dimensional non-sparse representation of the documents of every author [8]. Other studies even describe that not all words matter when establishing the profile of an author, but suggest that words near a personal pronoun are more discriminative for classifying an author's profile [10].

This year's author profiling shared task consisted of predicting gender and language variety of a group of authors in Twitter [13]. This group of Twitter users is distributed along four languages (English, Spanish, Portuguese and Arabic) and two genders (Male and Female). Furthermore, language variety depended on the respective language. English consisted of *Australia, Canada, Great Britain, Ireland, New Zealand* and *United States*. Spanish consisted of *Argentina, Chile, Colombia, Mexico, Peru, Spain* and *Venezuela*. Portuguese comprised *Brazil* and *Portugal*. Arabic included *Egypt, Gulf, Levantine* and *Maghrebi*. Unlike PAN AP 2016 shared task [14], PAN AP 2017 shared task uses the same domain of documents (Twitter) for training and testing [13].

Also, this is the first time language variety identification is added to the PAN AP shared task. Although language variety has been previously addressed as an author profiling task [12]. Furthermore, it has been part of the VarDial2016 (Workshop on NLP for Similar Languages, Varieties and Dialects) [9]. Malmasi et al. describe a big gap between traditional machine learning models and deep learning models in the participant teams evaluated in the VarDial2016. In this work we attempt to narrow this gap using convolutional neural networks (CNN) as a first approach for author profiling. CNNs have proven to be a successful method for classification of texts [6,5,17]. CNNs have also shown a good performance on authorship attribution tasks [16,15].

In this paper we describe our approach using CNNs for the author profiling task. CNNs are capable of capturing local-level interactions for learning profile-specific patterns. For every language, we trained separate models for gender and language variety using a CNN. Different CNN architectures were explored modifying parameters such as the size of input of the network, the size of the convolutional kernels, the number of kernels and the type of the input. We found experimentally that sequences of words performed better than sequences of characters as input for the CNN.

2 Methodology

In this section, we describe our submission to the PAN 2017 Author Profiling shared task. This architecture is an adaptation of previous work of CNNs at character level [16]. First, we briefly describe the preprocessing strategy, then how CNNs work, and finally we describe how CNN training was carried out. Authors' tweets were tokenized using a plain word tokenizer. We preserved both case and stopwords during preprocessing. After that all the tweets of an author are concatenated and split into k evenly sized sequences of texts. Words are then represented by non-sparse vectors of dimension e, also known as embeddings. As Figure 1 shows, a sequence of words is represented as a matrix $C \in \mathbb{R}^{e \times k}$ where each column corresponds to the word embedding vector value.

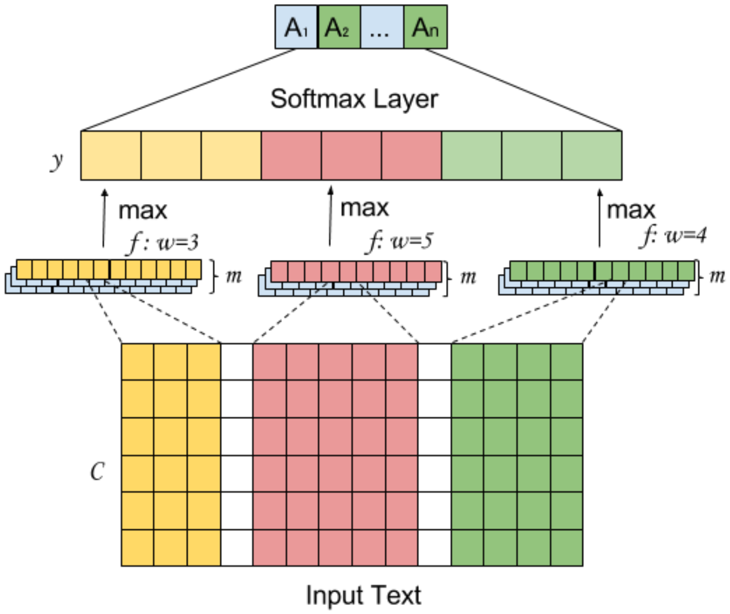

Figure 1. N-gram CNN. Word embeddings are fed to convolutional and max pooling layers, and the final classification is done via a softmax layer applied to the final text representation.

2.1 Word Convolutional Neural Networks

Word Convolutional Neural Networks (W-CNN) receive a fixed-length sequence of words as input. Figure 1 depicts the W-CNN architecture. W-CNN first layer applies a set of convolutional filters of different sizes. For the concrete case of Figure 1 $m = \{500, 500, 500\}$ and $w = \{2, 3, 4\}$. The convolution operation performed by these filters is only applied in one dimension. Then a max-pooling over time operation is performed over the output feature maps, where only the maximum value of each feature map is used. The max pooling outputs for each feature map are concatenated in a vector. Figure 1 shows the output vector of size 1500 composed by the maximum activation values generated by each convolutional filter over the input. Finally, a softmax layer is added, where its size A_n depends on the profiling task. Dropout regularization was also used after the Embedding layer with a $p = 0.25$. Given that we train our network using sequences of text of one author, we used a bagging scheme for prediction stage. If we have n sequences of text for one author, we generate n predictions for the corresponding author, then we average the predictions and get the class with the highest value. In that way an author is labeled with its respective gender and language variety.

2.2 Implementation details

Several CNN architectures were explored for finding the most suitable models for the author profiling task. Our exploration focused on two kinds of hyperparameters, Input-related and Convolution-related. For Input-related parameters we explored the type of input, the size of the input and the initialization values of the embeddings. The type of input were either tokenized sequences of words or sequences of char bigrams. The size of the input also was explored from a set of possible values $\{50, 100, 200, 300\}$. Larger input sizes mean a reduction in the number of training samples, making the training process difficult for complex architectures. Initialization values of the embeddings were also evaluated using either pretrained embeddings or embeddings trained from scratch using the supervised signal of the profiling task. Pretrained word embeddings were trained on Wikipedia for every language using FastText [2], although, we found that pretraining of the embeddings did not improve the results. For convolution-related parameters we explored the size w of the kernels and the number of kernels m. Larger size of kernels implies capturing long distance relationships between words, however this is only possible with a sufficient amount of training samples. Accordingly, we explored w from the set of values $\{1, 2, 3\}, \{2, 3, 4\}, \{4, 5, 6\}$, while the number of filters m varied from 1500 up to 3000. Also using a large number of filters, increases the representational capacity of the architecture, however it overfits quickly.

These architecture hyperparameters were found by exploration on the validation split of each setup and the best combination of parameters can be found in Table 1. We found also that word-based inputs performed better than char-based inputs over all the profiling setups. For training, we employed Keras [4]. We shuffled the samples into mini-batches of size 32 and used Gradient Descent with Adaptive Moment Estimation [7] with default learning rate. Validation loss was monitored during 100 epochs and only models with the best validation accuracy were saved and used for testing.

Layer	Parameters	English		Spanish		Portuguese		Arabic	
		Gender	Variety	Gender	Variety	Gender	Variety	Gender	Variety
Input	Input type	word	word	word	word	word	word	word	word
	Input size	200	200	300	200	200	200	50	300
	Pre-trained	No	No	No	No	No	No	No	No
Convolutional	m	1500	1500	1500	1500	1500	1500	1500	1500
	w	$[1, 2, 3]$	$[1, 2, 3]$	$[1, 2, 3]$	$[1, 2, 3]$	$[1, 2, 3]$	$[1, 2, 3]$	$[1, 2, 3]$	$[1, 2, 3]$

Table 1. Best combination of hyperparameters for the neural network architecture. Possible values for the hyperparameters are as follows: Input type can be word or char. Input size varied from $\{50, 100, 200, 300\}$. Pre-trained defined if embeddings were trained previously from Word2Vec or not. Convolutional number of filters m varied from $\{1500, 3000\}$. Convolutional sizes of filters w comprised $\{1, 2, 3\}, \{2, 3, 4\}, \{4, 5, 6\}$

.

3 Experiments and Results

This year's training data consists of 10800 Twitter users. For each individual author, an XML document is provided along with his/her tweets. There are 3000 documents for English, 4200 for Spanish, 1200 for Portuguese and 2400 for Arabic. For each language, we trained separately a model for gender and for language variety. For evaluation, we generated a stratified train/val split for every possible combination of **language_gender** and **language_variety**. Ten percent of the training documents was used for validation purposes.

The evaluation of the models for the shared task was performed using TIRA [11]. TIRA allows both organizers and participants to have a common framework for evaluation. Also, participants of a shared task can deploy and evaluate their method without accessing directly to the test dataset. We deployed on TIRA the best model found by validation. Table 2 shows the performance results of our method in the test dataset for the four languages. Accuracy is calculated separately for gender and language variation. For the joint column, accuracy is calculated on the basis that both gender and language variation were properly predicted.

Language	Joint	Gender	Variety
English	0.66	0.78	0.84
Spanish	0.73	0.77	0.94
Portuguese	0.81	0.82	0.98
Arabic	0.57	0.68	0.79

Table 2. Accuracy results on test dataset.

4 Discussion and Conclusion

Our architecture was evaluated over sequences of words and characters. We found experimentally better validation performances using word sequences. Although we also found that training a CNN for author profiling produces additional challenges such as hyperparameter tuning and quick overfitting. In our parameter exploration we encountered models that were prone to overfit at the very first epochs. We solved this introducing dropout regularization or using an architecture with a fewer number of parameters.

This work is a first approximation to the author profiling task using neural networks. Our system is capable of learning significant patterns without any handcrafted features, however it still performs worse than traditional methods that use a concatenation of content and style handcrafted features. Also, as it has been reported in previous works, content-based representations work better than style-based. In our future work, we will explore deeper convolutional networks with strong regularization, attention models and oversampling strategies. Also as suggested in [10], feeding sequences of text centered on personal pronouns could improve the performance of the CNN, because the network would only look at relevant examples.

References

1. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Commun. ACM 52(2), 119–123 (Feb 2009), http://doi.acm.org/10.1145/1461928.1461959

2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
3. Chambers, J., Trudgill, P., Schilling-Estes, N.: The Handbook of Language Variation and Change. Blackwell Handbooks in Linguistics, Wiley (2002), https://books.google.com.co/books?id=1ImVXjkmbHkC
4. Chollet, F., et al.: Keras. https://github.com/fchollet/keras (2015)
5. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 655–665. Association for Computational Linguistics, Baltimore, Maryland (June 2014), http://www.aclweb.org/anthology/P14-1062
6. 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. Association for Computational Linguistics, Doha, Qatar (October 2014), http://www.aclweb.org/anthology/D14-1181
7. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (2015)
8. López-Monroy, A.P., Montes-y Gómez, 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), http://www.sciencedirect.com/science/article/pii/S0950705115002427
9. 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 Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3). pp. 1–14. The COLING 2016 Organizing Committee, Osaka, Japan (December 2016), http://aclweb.org/anthology/W16-4801
10. Ortega-Mendoza, R.M., Franco-Arcega, A., López-Monroy, A.P., Montes-y Gómez, M.: I, me, mine: The role of personal phrases in author profiling. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 110–122. Springer International Publishing (2016)
11. 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: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
12. 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)
13. 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)
14. 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. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2016), http://ceur-ws.org/Vol-1609/
15. Ruder, S., Ghaffari, P., Breslin, J.G.: Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. arXiv preprint arXiv:1609.06686 (2016)

16. Shrestha, P., Sierra, S., Gonzalez, F., Montes, M., Rosso, P., Solorio, T.: Convolutional neural networks for authorship attribution of short texts. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 669–674. Association for Computational Linguistics, Valencia, Spain (April 2017), http://www.aclweb.org/anthology/E17-2106
17. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Advances in Neural Information Processing Systems. pp. 649–657 (2015)