Character-based Convolutional Neural Network and ResNet18 for Twitter Author Profiling

Notebook for PAN at CLEF 2018

Nils Schaetti

University of Neuchâtel
rue Emile Argand 11
2000 Neuchâtel, Switzerland
nils.schaetti@unine.ch

Abstract. This paper describes and evaluates a mixing model for multimodal author profiling using character-based Convolutional Neural Networks (CNN) for tweet classification and ResNet18 for images. We applied theses models to the author profiling task of the PAN18 challenge and show that its architecture allows these model to be applied to any language. For the tweets, a CNN based on an character-embedding layer, 1'500 filters and a temporal max-pooling layer reaches a classification accuracy of 74.87% with 100 tweets per author. For images, the ResNet18 model reaches a classification accuracy of 56.58% with 10 images per author. The evaluations are based on three collections of tweets and images (PAN AUTHOR PROFILING task at CLEF 2018).

1 Introduction

Today, web applications and social networks produce a large amount of data and various contents like pictures, videos and texts, shared directly from web sites and mobile devices. Social networks like Twitter are based on new kind of interactions with fast temporal dynamics generating an enormous variety of contents with their own characteristics which are difficult to compute with classical tools used on traditional texts like essays and articles.

We then face new questions : how to find differences in writing style on social networks between men and women, age groups, location or psychological profiles? The answers to these questions are important for the new problem we face in the era of social networks such as fake news, plagiarism, identity theft and astroturfing. Author profiling is then a problem of particular interest.

This paper is organised as follow. Section 2 introduces the dataset used for training, validation and testing, as well as the measures and methodology used to evaluate our approach. Section 3 explains the proposed character-based Convolutional Neural Network (CNN) model used to classify tweets. Section 4 describes the ResNet18 model used to classify images. In section 5, we evaluate the strategy we created and compare results on the three different test collections. In the last section, we draw conclusions on the main findings and possible future improvements.

2 Tweet Collections and Methodology

To compare different experimental results on the author profiling task with different models, we need a common ground composed of the same datasets and evaluation measures. In order to create this common ground, and to allow the large study in the domain of author profiling, the PAN CLEF [1] evaluation campaign was launched. Multiple research groups with different backgrounds from around the world have proposed a profiling algorithm to be evaluated in the PAN CLEF 2018 [8] [3] campaign with the same methodology [2].

All teams have used the *TIRA* platform to evaluate their strategy. This platform can be used to automatically deploy and evaluate a software [1]. The algorithms are evaluated on a common test dataset and with the same measures, but also on the base of the time need to produce the response. The access to this test dataset is restricted so that there is no data leakage to the participants during a software run. For the PAN CLEF 2018 evaluation campaign, three test collections of tweets and images were created, one for each of the following languages : English, Spanish and Arabic. Based on these collections, the problem to address was to predict the author's *gender* [4].

The training data was collected from Twitter. For each tweet collections, the texts come from the same language and are composed of tweets and images from authors, 100 tweets and 10 images per authors. For each author, there is a two-class label we can predict which can take the value *female* or *male*.

The test sets are also texts and images collected from Twitter and the task is therefore to predict the *gender* for each Twitter author in the test data. There is one collection per language (English, Spanish and Arabic). The English collection is composed of 3'000 authors, 1'500 for each gender, for a total of 300'000 tweets. The Spanish collection is also composed of 3'000 authors, 2'100 for each gender, for a total of 420'000 tweets. Finally, the Arabic collection is composed of 1'500 authors for a total of 150'000 tweets.

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

		Training			
Corpus		Authors	Tweets	Images	Gender
English	PAN17	3'600	360k	0	1'800; 1'800
	PAN18	3'000	300k	30k	1'500; 1'500
	Total	**6'000**	**660k**	**30k**	**3'300; 3'300**
Spanish	PAN17	4'200	420k	0	2'100; 2'100
	PAN18	3'000	300k	30k	1'500; 1'500
	Total	**7'200**	**720k**	**30k**	**3'600; 3'600**
Arabic	PAN17	2'400	240k	0	1'200; 1'200
	PAN18	1'500	150k	15k	750; 750
	Total	**3'900**	**390k**	**15k**	**1'950; 1'950**

Table 1: Corpora statistics

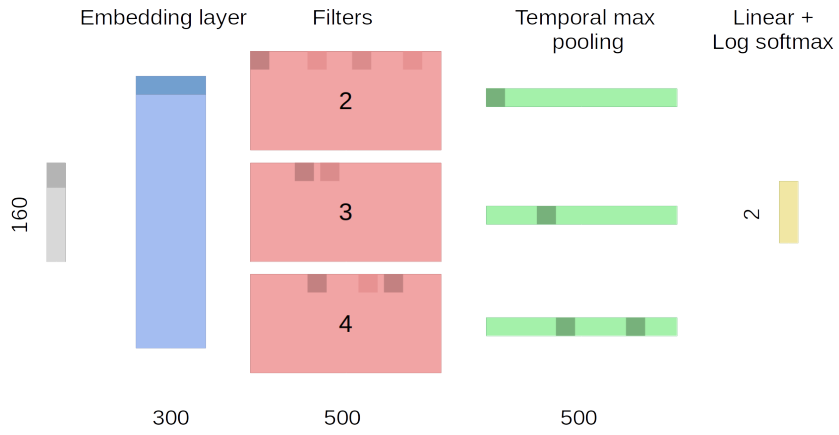

Fig. 1: Structure of the character 2-grams based *Convolutional Neural Network* with the following layers : embedding layer (dim=300), three convolutional layers (kernel size 2, 3 and 4), three temporal max pooling layers, a final linear layer of size 2 with log softmax outputs.

To allow our tweet classification model to reach high accuracy, we added the tweet collections of the author profiling task of the PAN17 to our training set. This result in a final training set of 6'000 authors for the English collection and a total 660'000 tweets. The final Spanish training set has 7'200 authors for 720'000 tweets and the final Arabic collection has 3'900 authors for a total of 390'000 tweets.

This year author profiling task proposes also to use 10 images from authors profile to build our model. There is 10 images per authors, resulting in 30'000 images for the English and Spanish collection, and 15'000 for the Arabic collection. An overview of these collections is depicted in table 1. The number of authors from the training set is given under the label "Authors" and the total number of tweets and images in the collection are indicated respectively under the labels "Tweets" and "Images". The label "Genders" indicates the number of authors for each gender. The training data set is well balanced as for each collection, there is the same number of authors for each gender. The Spanish collection is the biggest with 7'200 authors, and the smallest is the Arabic collection with 3'900 authors.

A similar test set will be used to compare the participants' strategies of the PAN CLEF 2018 campaign, and we don't have information about its size due to the *TIRA* system. The response for the gender is a binary choice (*male / female*). The overall performance of the system is the joint classification accuracy for both tweets and images, but is also evaluated for each specific task. The accuracy is the number of authors where the gender is correctly predicted for the same author divided by the number of authors in the collection.

3 Character-based Convolutional Neural Network (CNN)

In machine learning, a *Convolutional Neural Network* (or CNN) is a kind of feed-forward artificial neural network, in which the patterns of connection between the neu-

rons are inspired from the visual cortex.

In our system, we applied a character 2-grams based CNN to each tweet in a collection. A tweet is fed into the model as an array of character 2-grams with a fixed size of 160. If the tweet is shorter than 160, the additional space is felt with zeros as each character 2-grams is represented by an index. For each tweet, we removed all URLs and we passed it to lower cases, we then transformed it into a list of overlapping character 2-grams. Each character 2-grams are transformed to indexes with a vocabulary constructed during the training phase and character 2-grams which appear in the test set but not in the training set are set to zero.

The first layer is a *embedding layer* with a size equal to the vocabulary size and a dimension of 300 for each character 2-grams. This layer has two purposes, first to reduce the dimensionality of the inputs to 300, compared to $|V|$ for one-hot encoded vectors, and secondly, to encode similarities between character 2-grams into a multi-dimensional space where two character 2-grams appearing in similar context are near each others.

The second layer is composed of three different *convolutional layers* with kernel sizes of 2, 3 an 4 followed by a *ReLU* nonlinearity. Theses layers encode patterns of 2, 3 or 4 consecutive character 2-grams and each layer has 500 filters and 1500 patterns can thus be represented. During the training phase, our model will then find the 1500 most effective patterns of character 2-grams that allow the two classes to be separated.

The third layer is composed of three *temporal max pooling layers*, one for each preceding convolutional layers. The temporal max pooling meaning here that these layers take the maximum of the each preceding filters on the entire input. We concatenate the tree output and end up with a vector of size 1500 representing at what magnitude each filter appears in the input tweet.

The final layer is a *linear input* of size 2, one output per class, followed by a log-softmax so that the outputs is the probability for each class. This layer encodes the combination of filter matches representing each class. The training phase consists of using the PAN17 and PAN18 for training. We used the *stochastic gradient descent algorithm* to train our model with a *learning rate* of 0.001, a *momentum* of 0.9 and *cross-entropy* as *loss function*. We trained our model for 150 iterations.

To implement our model, we used TorchLanguage [6], a package based on pyTorch designed for Natural Language Processing.

4 ResNet18 for Image Gender Classification

A *ResNet* is a kind of deep neural network using *skip connections* or *short-cuts* which jumps over some layers. ResNet models were introduced in 2015 and won several competition in computer vision.

ResNet has structures similar to the brain's cerebral cortex, however it is not clear how many layers there is in the cerebral cortex compare to layers in ResNet and if all area of exhibits the same structure, but they look quite similar over large areas. The motivation behind skipping layers in ANN is to avoid the well known problem of vanishing gradients using activation from a previous layer until the next one has learned its weights. Figure 2 shows ResNet's basic building block.

To implement this model, we used a pre-trained model available with TorchVision, which proposes such model with 18, 34, 50, 101 and 152 layers. We choose the model with 18 layers to avoid overfitting as the training set is not very large.

In the training phase, we used 90% of the PAN18 image training set for training and 10% to evaluate the performances at each iteration. We used the *stochastic gradient descent algorithm* to train our model with a *learning rate* of 0.001, a *momentum* of 0.9 and *cross-entropy* as *loss function*. At the end of the training phase, we choose the ResNet18 which obtained the best accuracy on the validation dataset. We trained the model on the whole training set independently of the language, we end up then with a single model for all collections.

5 Evaluation

As we have two models to predict the class of a single instance, we needed to define how to compute the final decision for a profile. For the two models, the decision for a profile is the average class probabilities over all instances, 100 for the tweets, 10 for the images. The final decision is based only on tweets as all tests to take also images into account was lowering the overall results.

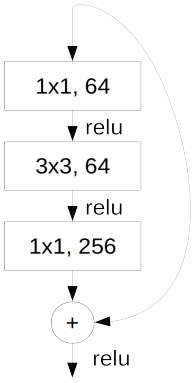

Fig. 2: Example of building blocks for ResNet18.

The table 2 shows the results on the three test collections obtained on the *TIRA* platform. The Spanish is the hardest to profile with an overall accuracy of 73.59% on tweets and 57.63% on images. For the English language collection, the accuracy goes from 77.11% on tweets to 57.82% on images. Finally, the accuracy on the Arabic language collection goes from 73.90% on tweets to 54.30% on images. The Arabic image collection is the harder to predict with 54.30% against 57.82% and 57.63 respectively for the English and Spanish collection. On the other hand, the Spanish tweet collection is the harder to predict with 73.59% against 73.90% and 77.11% respectively for the Arabic and English collections.
These results show an improvement compared to our last year results on the gender profiling task (PAN CLEF 2017) where we used a CNN model based on a matrix of character bigrams [5, 7].

6 Conclusion

This paper proposes a combination of Deep-Learning model for Twitter user profiling based on 100 tweets and 10 images per author. Based on the hypothesis that an author's writing style and the images posted on a social networks can be used to extract its gender, we introduced a CNN classifier for tweet classification and a ResNet18 model for image classification that can effectively predict this characteristics. The character 2-grams based CNN shows a good language-independent performance on tweet gender classification, on the other hand, the ResNet18 model is effective on image classification for gender profiling.

Corpus	Gender	Images	**Both**	*Random*
English	0.7711	0.5782	**0.7711**	0.5000
Spanish	0.7359	0.5763	**0.7359**	0.5000
Arabic	0.7390	0.5430	**0.7390**	0.5000
Overall	0.7487	0.5658	**0.7487**	0.5000

Table 2: Evaluation for the three *test* collections

The CNN model achieves its best performance on the test dataset on the English collection with 77.11% accuracy, and a very good accuracy of 73.90% and 73.59% on the Arabic and Spanish collection respectively. The biggest challenge of this year's PAN author profiling task were the gender classification problem based on images posted by the user where our model achieve an average of 56.58%.

References

1. 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)
2. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Jones, G., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, Septembre 11-14, 2017, Proceedings. Springer, Berlin Heidelberg New York (Sep 2017)
3. 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)
4. Rangel, F., Rosso, P., Potthast, M., Stein, B.: In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Labs Working Notes
5. Schaetti, N.: Unine at clef 2017: Tf-idf and deep-learning for author profiling. PAN CLEF 2017 (2017)
6. Schaetti, N.: Torchlanguage: Natural language processing with pytorch. https://github.com/nschaetti/TorchLanguage (2018)
7. Schaetti, N., Savoy, J.: Comparison of Neural Models for Gender Profiling. In: Domenica Fioredistella Iezzi, Livia Celardo, M.M. (ed.) Proceedings of the 14th international conference on statistical analysis of textual data (Jun 2018)
8. 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)