Gender Prediction From Tweets With Convolutional Neural Networks

Notebook for PAN at CLEF 2018

Erhan Sezerer, Ozan Polatbilek, Özge Sevgili, and Selma Tekir

Izmir Institute of Technology
{erhansezerer, ozanpolatbilek, ozgesevgili, selmatekir}@iyte.edu.tr

Abstract This paper presents a system[1] developed for the author profiling task of PAN at CLEF 2018 . The system utilizes style-based features to predict the gender information from the given tweets of each user. These features are automatically extracted by Convolutional Neural Networks (CNN). The system mainly depends on the idea that the informativeness of each tweet is not the same in terms of the gender of a user. Thus, the attention mechanism is included to the CNN outputs in order to discriminate the tweets carrying more information. Our architecture was able to obtain competitive results on three languages provided by the PAN 2018 author profiling challenge with an average accuracy of 75.1% on local runs and 70.23% on the submission run.

1 Introduction

Author profiling is the characterization of an author through some key attributes such as gender, age, and language. It's an indispensable task especially in security, forensics, and marketing. In the security world, predictive profiling is a measure for proactive threat assessment. In forensics; profiling is used to support attribution for an incident, while in marketing it helps to prepare targeted advertisements.

In today's social media-driven environment, automatic user profiling is not the same as before because what the users write and share in social media provide a great data source for the potential learning approaches. As a general rule, more data make classifiers more accurate.

In more technical terms, author profiling is defined as a classification task where the aim is to predict the attribute of an author out of the given attribute classes. The traditional machine learning process is followed to fulfill the task. Feature selection is an important part of the process. Literature categorize the types of features that can be used for authorship profiling as content-based features and style-based features. Evidence proved that the most effective style-based features for gender discrimination are determiners and prepositions (markers of male writing) and pronouns (markers of female writing). As for content-based features, words related to technology (male) and words related to personal life or relationships (female) are proved to be most useful [1].

[1] The implementation can be found at: https://github.com/Darg-Iztech/Gender_Classification

The recent deep learning-based approaches take prominence in this area as they perform feature selection automatically. We tackled the problem in a similar way. The proposed approach feeds the characters of a specific user's tweets into the system, where the system learns the embeddings character to character and it runs a Convolutional Neural Network (CNN) for each individual tweet of the user. Then, CNN outputs are combined and pass through an attention layer to form the user specific vector for prediction.

In this work, we aim to obtain style-based features from the tweets of users by using CNNs. CNNs are known to be good at identifying the local patterns from the inputs [5]. They were originally designed to tackle the problems in vision tasks by identifying the small objects or patterns in images [9], but later, they were introduced into NLP tasks to extract the syntactic, local features from the text [4].

In PAN 2018 [16] author profiling task [15], the profiling dimensions are determined as gender and language, where the selected languages are English, Spanish, and Arabic respectively. As for training data; in addition to text in the form of tweets, the user shared images are provided as well. Thus, hybrid solutions that use both text and image-based features are encouraged.

Our system uses only text-based features. The basic characteristics of our approach can be highlighted as follows:

- The system learns on a user basis iteratively.
- The input is in the form of characters.
- A CNN per-tweet is constructed to identify local tweet-wide indicators in larger user profile vector.
- An attention layer is used to combine CNN outputs using normalized weights.

In the remaining part of the paper, we first present the related work. In Section 3, the proposed method is explained in detail. Then, the performance is tabulated and evaluated. Finally, in Section 5, the paper is concluded with some remarks and possible future directions.

2 Related Work

Argamon et al. [1] categorize the types of features that can be used for authorship profiling as content-based features and style-based features. Their experiments show that the most effective style-based features for gender discrimination are determiners and prepositions (markers of male writing) and pronouns (markers of female writing). As for content-based features, words related to technology (male) and words related to personal life or relationships (female) are proved to be most useful.

Rangel and Rosso [14] investigate the impact of emotions in age and gender identification. They process text, create part-of-speech (POS) tag graphs (POS tags as nodes and their sequence as edges) and expand those graphs by related topic words, polarity labels, and emotion words from the emotion dictionary. Then, they extract features using graph analytics and feed them into machine learning algorithms to make the classification. Their results prove that language use and emotions are effective in discriminating gender and age.

In the overview paper of the Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter [12], the participant systems are compared with respect to features and classification approaches. In that edition of the author profiling task, more participants employ deep learning techniques, which perform automatic feature selection. In the gender and language variety subtasks; the best performances belong to a logistic regression classifier with combinations of character, word, and POS n-grams, emojis, sentiments, character flooding, an SVM trained with combinations of character and tf-idf n-grams, and a deep learning approach combining word and character embeddings with CNN, RNN, attention mechanism, max-pooling layer, and fully-connected layer.

Basile et al. [3] try a Support Vector Machine (SVM) with word unigram and character n-grams on PAN 2017 author profiling task where they have best results among other competitors. They use character three to five grams and word uni to bi-grams with tf-idf weighting and use SVM on this feature space to discriminate both gender and language variety. They also mention that the hand-crafted features decrease accuracy rather than helping in this specific task.

Miura et al. [11] propose two deep-learning based approaches which combine both context-based and style-based features by taking the word level and the character level information of the tweet's text. Their systems use both Recurrent Neural Network (RNN) (to address context-based features with the given word information) and CNN (to address style-based features with the given character information). Their architectures consist of attention mechanism layers, a max-pooling layer, and also fully-connected layers. The difference between the architectures is that one of them is on a tweet-basis while the other one is on a user-basis. Additionally, the places of layers lead to another difference.

Kodiyan et al. [8] also use a deep learning approach by implementing a bidirectional RNN with Gated Recurrent Units. They add an attention layer on tweet level to learn the most important parts of each tweet. In order to move from tweet level to user level they add the tweet predictions of a user together and use it as a single user level prediction.

3 Method

In this section, the description of the dataset and the details of the proposed model are given including choice of parameters, preprocessing steps and architectural details.

3.1 Data

PAN 2018 Author Profiling dataset [15] is based on 3 languages (English, Arabic, Spanish) with ground-truth gender information. It has 3000 users for English, 3000 users for Spanish, and 1500 users for Arabic language where each user has 100 tweets and 10 images that they posted on Twitter. In this work, only text data are used in gender classification.

3.2 Preprocessing

In Twitter, characters are used not only to create words but also to express emotions like smiling as ':)' or blinking as ';)', because of this type of usage, punctuations and stop words did not get eliminated, texts are given as how they are. NLTK [10] is used to tokenize tweets. To illustrate (example from NLTK):

Tweet = "This is a cooool #dummysmiley: :-) :-P <3 and some arrows < > -> <–"

Tokenized Tweet = ['This', 'is', 'a', 'cooool', '#dummysmiley', ':', ':-)', ':-P', '<3', 'and', 'some', 'arrows', '<', '>', '->', '<–']

Each word in the tokenized tweet is applied lowercasing. Then, each character from the word is taken to be utilized in the input to the system. Thus, the tweet in the above example is turned into the following input:

Input = ['t', 'h', 'i', 's', 'i', 's', 'a', 'c', 'o', 'o', 'o', 'o', 'l', '#', 'd', 'u', 'm', 'm', 'y', 's', 'm', 'i', 'l', 'e', 'y', ':', ':', '-', ')', ':', '-', 'p', '<', '3', 'a', 'n', 'd', 's', 'o', 'm', e', 'a', 'r', 'r', 'o', 'w', 's', '<', '>', '-', '>', '<', '-', '-']

For each user, the number of characters is set to the highest number that is allowed for tweets in Twitter. If a tweet has fewer number of characters than the maximum, padding is applied to the end of the tweet.

3.3 Character Embeddings

Character embeddings with size 25 are initialized by sampling from uniform distribution with 0 mean and trained simultaneously with the neural network. Due to their smaller size and count, training character embeddings requires fewer text to be trained than word embeddings. Therefore, the given dataset was sufficient to train them and no additional data are collected or used.

3.4 Architecture

In this study, each tweet of a user is passed to the CNN simultaneously as a sequence of characters to assess the style-based features of each particular tweet. CNN outputs a feature vector for each tweet.

At this level, using other methods like combining, flattening or averaging the feature vectors would mean to explicitly assume the equal importance among tweets. However, the level of information on gender may differ from tweet to tweet. Therefore, A Bahdanau attention mechanism [2] is combined with the character CNN in order to learn which tweet holds more information on the gender of its author. Figure 1 shows the attention mechanism in detail which is calculated by the following formulas:

$$A_i = tanh(\boldsymbol{W_\alpha} t_i + b) \tag{1}$$

$$v_i = \frac{exp(A_i w_i)}{\sum_j exp(A_j w_j)} \tag{2}$$

$$o_i = v_i t_i \tag{3}$$

$$K = \sum_i o_i \tag{4}$$

where $\boldsymbol{W_\alpha}$ is a weight matrix used to multiply each output of the CNN, t_i is the ith tweet, b is bias vector, w_i is the attention weights, A_i is the attention context vector, v_i is the attention value for ith tweet, o_i is attention output vector for the corresponding tweet, K is the output vector for user.

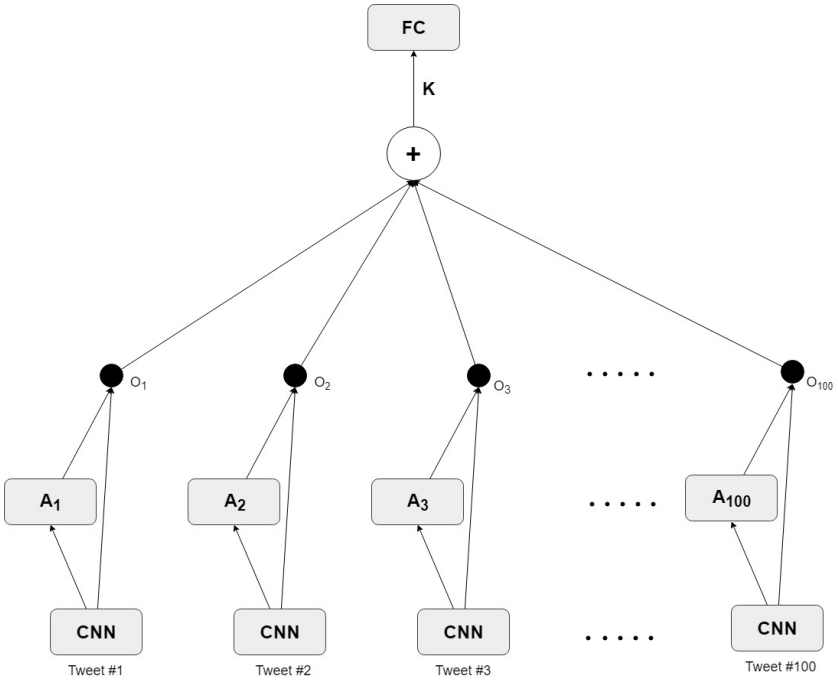

Figure 1. Attention mechanism.

A fully connected layer is used on the output of the attention layer to reduce the size of the feature vector to the number of genders. Predictions are obtained after applying softmax over the output of the fully connected layer. Proposed model can be seen in Figure 2.

CNN [6][2] is implemented with ReLu activation function and [filter size, embedding size] shaped filters with stride 1 to make all characters visited. Adam optimizer [7] is used with cross entropy loss. To prevent the model from overfitting L2 regularization loss is used.

[2] implementation can be found at: https://github.com/dennybritz/cnn-text-classification-tf

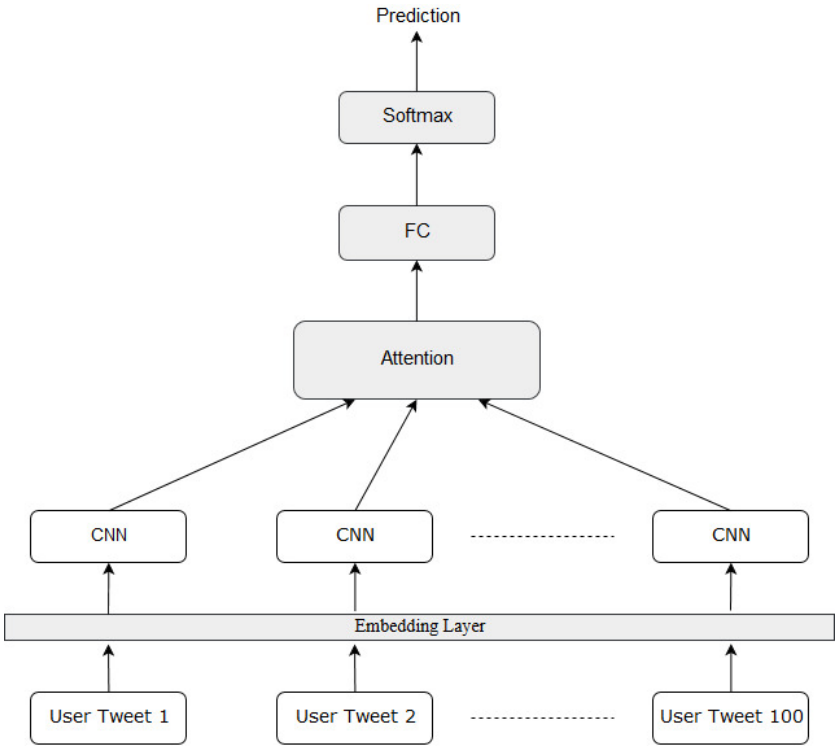

Figure 2. The proposed model.

3.5 Parameter Selection

Exhaustive grid search is used to optimize the hyperparameters of the model. Parameters we have tried for each language can be seen in Table 1. Due to differences in each language and the size of the dataset, different hyperparameters gave best results for each language (Table 2) .

Table 1. Hyperparameters used in optimizations.

Parameter	Values
Embedding Size	25
Learning Rate	$5x10^{-3}$, 10^{-4}, $5x10^{-4}$, 10^{-5}, $5x10^{-5}$, 10^{-6}
L2 Regularization Coefficient	$5x10^{-4}$, 10^{-5}, $5x10^{-5}$, 10^{-6}, $5x10^{-6}$, 10^{-7}, $5x10^{-7}$, 10^{-8}
Filter sizes	3, 5, 6, 9
Number of Filters	40, 50, 60, 75, 100

Table 2. Parameters with tuned values.

Parameter	English	Spanish	Arabic
Embedding Size	25	25	25
Learning Rate	10^{-4}	10^{-4}	10^{-4}
L2 Regularization Coefficient	10^{-6}	5x10−6	10^{-6}
Filter sizes	3, 6	3, 6	3, 6, 9
Number of Filters	75	60	50
Strides	1	1	1

4 Results

We have selected the model with the best working parameters, shown in Table 2. As can be seen from Table 3, our best model gives between 70% and 79% accuracy for different languages in our validation runs. In the submission run over TIRA framework [13], our best models obtained approximately 4% lower accuracy than the validation runs for each language.

Table 3. Gender prediction accuracy for each language.

Language	Validation Accuracy(%)	Test Accuracy(%)
English	79.0	74.95
Arabic	75.7	69.20
Spanish	70.7	66.55
Average	75.1	70.23

Table 4. Accuracy(%) of models with and without attention mechanism

Language	CNN without attention	CNN with attention
English	76.3	79.0
Arabic	72.0	75.7
Spanish	66.3	70.7
Average	71.5	75.1

We have also observed in our experiments that, instead of averaging the feature vectors at the output of the CNN or using fully connected layers to combine them, using attention increases the accuracy of the system by approximately 3 percent on an average in three aforementioned languages (Table 4). This shows that the attention layer was able to learn "where to look" and identify the tweets that are more informative when it comes to gender prediction. Table 5 shows an example of attention values for three tweets of a particular user for each gender where the attention values correspond to the probabilities of the respective tweets over the hundred tweets provided for the user by the PAN author profiling dataset. It can be seen that the attention layer was

able to assign higher values to tweets which have stronger gender indicators such as the words "bro" for male, "love" for female whereas it assigned low scores to automatically generated tweets like the third tweet of the male user.

Table 5. Example of attention values on tweets for two users

User	Tweets	Attention values
Male	@******* bro it's 1 sub	0.04344852
	Recorded 2 videos today and I think they are going to be my best ever videos ever! Not sure when I will upload it tho :/	0.01365168
	Welcome to my new 9 followers and goodbye to 3 unfollowers (FREE stats by https:*******	0.00081017
Female	Love hearing from the women themselves about their own experiences. including from the trans community. https:*******	0.12932625
	I have to declare I love the Great British Sewing New..for reality TV they are such kind generous people	0.01901261
	If you caught Prt 2 of "The Oldest Profession" on Night,s hear the other 2 progs: https:******* or download all 3 @*******	0.00156761

5 Conclusion

We have described a system submitted to the author profiling task of PAN at CLEF 2018. A CNN architecture is proposed which takes the characters of each tweet's text as an input. This input is based on a user in which each tweet of the user given to the system. Local style-based features are aimed to be extracted by this system, automatically. The critical issue related with the proposed system is to recognize that each tweet can carry different level of information to discriminate the gender of a user. The attention mechanism is able to catch that difference. This mechanism is added to CNN outputs. Therefore, the predictions are based on the tweets holding more information about the gender. As an output, the system gives the prediction of user's gender in a vector form.

In the given dataset, in addition to tweets, there are images posted by the users. In future, we are also planning to make use of those image data along with our current architecture and we are expecting to get improved results due to that addition.

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. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations (2014), http://arxiv.org/abs/1409.0473
3. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-gram: New groningen author-profiling model. CoRR abs/1707.03764 (2017), http://arxiv.org/abs/1707.03764

4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. 12, 2493–2537 (Nov 2011), http://dl.acm.org/citation.cfm?id=1953048.2078186
5. Goldberg, Y., Hirst, G.: Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers (2017)
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 (2014)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR abs/1412.6980 (2014), http://arxiv.org/abs/1412.6980
8. Kodiyan, D., Hardegger, F., Neuhaus, S., Cieliebak, M.: Author profiling with bidirectional rnns using attention with grus. In: CLEF (2017)
9. Lecun, Y., Bengio, Y.: Convolutional networks for images, speech, and time-series. In: Arbib, M. (ed.) The handbook of brain theory and neural networks. MIT Press (1995)
10. Loper, E., Bird, S.: Nltk: The natural language toolkit. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1. pp. 63–70. ETMTNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
11. Miura, Y., Taniguchi, T., Taniguchi, M., Ohkuma, T.: Author Profiling with Word+Character Neural Attention Network—Notebook for PAN at CLEF 2017. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland. CEUR-WS.org (Sep 2017), http://ceur-ws.org/Vol-1866/
12. Pardo, F.M.R., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. In: CLEF (2017)
13. 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)
14. Rangel, F., Rosso, P.: On the impact of emotions on author profiling. Information Processing & Management 52(1), 73 – 92 (2016), http://www.sciencedirect.com/science/article/pii/S0306457315000783, emotion and Sentiment in Social and Expressive Media
15. 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)
16. 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)