Twitter Text and Image Gender Classification with a Logistic Regression N-gram Model

Notebook for PAN at CLEF 2018

Moniek Nieuwenhuis and Jeroen Wilkens

University of Groningen, The Netherlands
{m.l.nieuwenhuis,j.r.wilkens}@student.rug.nl

Abstract We present our participation in the PAN 2018 Author Profiling shared task, classifying authors on gender for English, Arabic and Spanish. We participated in all sub-tasks and propose a system for classification with text, images and the combination of those two. Our final submitted system is a Logistic Regression classifier that uses word and character n-grams as textual features and a set of automatically derived image-based features such as the presence, proportion and number of faces to detect selfies as well as the faces' emotions and gender. We experimented with word embeddings, which negatively affected our system's performance. Our cross-validated training results shows slight improvements in performance for Arabic and Spanish when image-based features are added to text-based features. Our highest scores on the PAN 2018 test dataset are accuracies of 81.2% for English using only text-based features, 78.7% for Arabic using both text- and image-based features and 80.3% for Spanish using only text-based features. Overall, we finished 6th in the global ranking with an average accuracy for our text and image combination system of 79.6%.

1 Introduction

The field of author profiling is about inferring traits from an author such as gender, age and personality. With the rise of social media platforms, such as Twitter and Facebook, the field of author profiling has gained more interest. From multiple viewpoints, it is desirable to profile an author. Examples of such viewpoints could be from a security point of view in order to detect authors with criminal intentions and from a marketing point of view in order to narrow down target audiences for online advertisements.

In the past years, multiple shared tasks have been organized on the topic of author profiling [18,16]. In this paper, we describe our approach for the Author Profiling shared task at PAN 2018 [17]. This year's Author Profiling task, is the 6th iteration of this task and is slightly different from the previous years, since the gold standard data now includes images. The task is to built a system to classify a Twitter author's gender by 100 of its tweets and 10 posted images. Though the images are new for this shared task, previous work already created systems that are capable of detecting gender, emotional expressions and personalities from images [2]. By combining such image classification systems with textual classification systems, it can be determined whether this addition of images can improve the final accuracy.

In the last two years, the winning systems for 2016 [22] and 2017 [3] were both SVM classifiers that made use of word n-grams and character n-grams. Although deep-learning methods were introduced, such as Recurrent Neural Networks [7] and Convolutional Neural Networks [19,20], they haven't been able to beat those systems yet. Therefore, our approach will focus on the successful models of the previous iterations of the shared task, a SVM classifier such as in [3] and a Logistic Regression classifier as used in [10], in which we will take these systems as baselines and try to improve them by performing a parameter search, experimenting with word embeddings and adding image-based features. The latter includes a feature to indicate selfies, since females tend to post more selfies than males [5,21].

2 Method

2.1 Data

The PAN 2018 training corpus consists of tweets from three different languages, English, Spanish and Arabic. For each author there are 100 tweets and 10 images labeled by gender. The gender labels (male and female) are evenly distributed over the training corpus. Table 1 shows an overview of the PAN 2018 training corpus released by the organization.

Table 1. PAN 2018 dataset overview.

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

2.2 N-grams

The main set of features we used were n-grams. The winners of the previous year's Author Profiling shared task [3], as well as [1,4,6,9,10,12,13,19] showed that word n-grams and character n-grams are very robust features for this task. Another advantage of using n-grams is that they are non-handcrafted features and thus easy to generate. Also, there is no dependence on either pre-trained word embeddings, or large corpora of text to train word embeddings. For all three languages, we experimented with different lengths of word and character n-grams.

2.3 Word Embeddings

Aside from the n-gram features, we experimented with using word embeddings. For English, we experimented with pre-trained word embeddings from GloVe [15]. We used

embeddings with vector lengths of 100 and 200 dimensions that were created from a corpus consisting of 2 billion tweets containing 27 billion tokens. For Spanish, we used pre-trained word embeddings from [8] with vector lengths of 200 dimensions. These embeddings were constructed from a total amount of 58.7 million Spanish tweets having 1.1 billion tokens. For Arabic, we trained our own word embeddings from roughly 70 million recently scraped Arabic tweets with vector lengths of 200 dimensions.

2.4 Images

This year's new addition to the gender classification task is classification by images. Our approach to use the images for this task is to utilize existing image feature extraction tools from related research. In our system, we have used the software used in [2].

In that study, a convolutional neural network (CNN) was implemented to find and classify faces in images by gender and emotion. The CNN model contains of 4 residual depth-wise separable convolutions, whereby each convolution is followed by a batch normalization operation and a ReLU activation function. The last layer of the model applies a global average pooling and soft-max activation function to produce the prediction. The system achieved an accuracy of 95% on the IMDB gender dataset and 66% on the FER-2013 emotion dataset. That system, including all code and pre-trained models are available under an open-source license.[1]

The software from [2] was implemented in our system without preprocessing the images. The software converts the images from a Twitter user to a set of 13 features:

1. Average number of faces
2. Number of images that include a face
3. Average area the faces take up
4. Average area the largest face take up
5. Average number of men
6. Average number of women
7. Percentage of faces being angry
8. Percentage of faces being disgusted
9. Percentage of faces being fearful
10. Percentage of faces being happy
11. Percentage of faces being sad
12. Percentage of faces being surprised
13. Percentage of faces being neutral

The first two features are about the presence and number of detected faces in the images. In Table 2 are the number of faces detected for each language and gender. We see that there is little to no difference between the genders for which a face is detected.

Features three and four cover the relative area of images that are covered by a detected face. These features are intended to capture selfies. Previous research [5,21] studied selfie-related behaviours between males and females. One of the findings from those studies is that females tend to make and post more selfies on social media. Having a

[1] `https://github.com/oarriaga/face_classification`

Table 2. Amount of users for which a face is detected.

	Male	Female
English	1.401	1.396
Arabic	683	651
Spanish	1.430	1.394

large face area on an image with only one detected face could identify that an image is a selfie, and therefore these features could be helpful in the classification task.

Features five and six are about the gender of the detected faces. The values of these two features are floats ranging from 0 to 1, indicating the proportion of each gender. In Table 3 are the probabilities for a gender posting more faces of a specific gender in a image. We see that for all languages, males are posting more images of male faces than female faces, especially for Arabic there is a large difference in male and female faces. English and Spanish females are slightly posting more images with female faces than male faces, except for Arabic, in which female user post more male faces than female faces, but still to a lesser extent compared to their male counterparts.

Table 3. The probability that a face in a image is a male or female per gender.

	Male		**Female**	
	Male faces	**Female Faces**	**Male Faces**	**Female Faces**
English	0.598	0.402	0.441	0.559
Arabic	0.740	0.260	0.564	0.436
Spanish	0.622	0.378	0.496	0.504

Lastly, when one of the seven emotions from [2] could be detected, which the software was not always capable of, we stored the proportion of these emotions in seven float values ranging from 0 to 1. Table 4 shows an overview of the emotions for each gender and language. The table shows that, generally, there are small to no differences between male and female regarding emotions. The only conclusion that holds for all languages is that females tend to post more happy people. For English, males post more images of angry people, but for Arabic this is the opposite. Also, no one is ever surprised, raising the question whether the system of [2] can accurately detect this. Overall, we expect that these features will not be (very) beneficial for our system.

2.5 Models

To get to our best system, we experimented with different classifiers. We used the Python package Sklearn [14] to implement the LinearSVC classifier as used in [3] and the Logistic Regression classifier with the parameters C = 1e2 and fit_intercept = False as used in [10], we also tried a K-Nearest Neighbour classifier.

Table 4. Emotion probabilities of detected faces per language and gender.

		Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
English	**Male**	0.061	0.001	0.030	0.280	0.098	0.000	0.167
	Female	0.049	0.001	0.036	0.330	0.107	0.000	0.148
Arabic	**Male**	0.068	0.002	0.034	0.212	0.119	0.000	0.155
	Female	0.075	0.001	0.028	0.255	0.152	0.000	0.206
Spanish	**Male**	0.058	0.001	0.026	0.217	0.110	0.000	0.179
	Female	0.053	0.004	0.027	0.260	0.101	0.000	0.174

The results of all tested classification models can be found in Table 5. For every model,we measured its performance by accuracy in a 10-fold cross-validation setup. The models are all using the n-gram features used as in [3]. We found that using the Logistic Regression classifier resulted in the best performance, meaning we will use this classifier for our next experiments.

Table 5. Results on text of different models on 10-fold CV.

System	**En**	**Ar**	**Es**
SVM	0.826	0.772	0.773
Logistic regression	0.831	0.779	0.776
K-Nearest Neighbour	0.647	0.622	0.597

For the logistic Regression model we performed a parameter search, mainly to find the optimal number of word and character n-grams. Our baseline n-gram model was the n-gram model used in [3], which was using word 1- and 2-grams and character 3- to 5-grams. We tested different settings of n-grams but we were unable to outperform the settings from [3]. Table 6 shows the results for the best settings, as well as the best results found for word and character n-grams apart.

Table 6. Results of different n-gram combinations for the Logistic Regression model (10-fold CV).

N-grams	**En**	**Ar**	**Es**
word n-grams (n=1,2) + char n-grams (n=3,4,5)	0.831	0.779	0.776
bag of words	0.811	0.769	0.767
word n-grams (n=1,2)	0.804	0.757	0.756
char n-grams (n=3,4,5)	0.814	0.789	0.776

2.6 Text Preprocessing

For the preprocessing of the text data we lowercased all tweets and subsequently tokenized the tweets with the NLTK Tweet Tokenizer.[2] We also replaced every username with `@username` and every URL to `URL`.

Table 7 shows that our preprocessing methods do indeed improve performance for each language. Especially generalizing over URLs was beneficial.

Table 7. Accuracies of adding the different preprocessing methods, using 10-fold CV.

Preprocessing	**En**	**Ar**	**Es**
Baseline	0.816	0.754	0.759
+ Tokenization	0.818	0.764	0.760
+ Lowercasing tweets	0.818	0.764	0.760
+ URL to `URL`	0.827	0.774	0.767
+ Usernames to `@username`	0.831	0.779	0.776

3 Results

In this section we will report the results of our systems on the training corpus (10-fold CV) and final test set.

3.1 Training Results

The results in Table 8 shows that only using n-grams as features results in having a good baseline result that is in line with the findings in [3]. The model that only uses embeddings performs worse than the model that utilizes n-grams. Moreover, the model that combines embeddings and n-grams also performs worse.

The image-only model performed poorly with accuracies around 60%. However, using these features in combination with the n-gram features gave us slight performance improvements for Arabic and Spanish. Although we have found an increase of accuracy in Arabic, approximate randomization testing [11][3] showed us that this improvement is not significant.

3.2 Official Results

We handed in three final systems; one for classification based on text-only, one for images-only and one for the combination of text and images. For all of the three final systems, we used a Logistic Regression classifier.

[2] `http://www.nltk.org/`

[3] We used the script by Vincent Van Asch `https://www.clips.uantwerpen.be/scripts/art`

Table 8. Accuracy of features with logistic regression model average of 10-fold CV.

Features	En	Ar	Es
n-grams	0.831	0.779	0.776
embeddings	0.786	0.725	0.759
images	0.604	0.618	0.592
n-grams + embeddings	0.775	0.728	0.755
n-grams + images	0.823	0.792	0.778
n-grams + embeddings + images	0.815	0.715	0.741

Table 9 shows our official results on the PAN 2018 test set. We see that our performance for English is a bit lower than we expected, but for Spanish and Arabic we obtain a better performance than on the training set. On average over all languages we scored an accuracy of 0.799 on the text-only submission. The images do not influence the score much, but since it now decreases the score, it is questionable whether our (simplistic) approach of processing the images is helpful for this task.

For the combination system with both text and image features we scored an average accuracy score of 0.7963 and became 6th in the global ranking. In general we can say that our Logistic Regression model does quite well, making it a robust, straightforward and reliable method of doing gender classification.

Table 9. Results on PAN 2018 test set.

Language	Text	Image	Combination
English	0.812	0.610	0.810
Arabic	0.783	0.623	0.787
Spanish	0.803	0.587	0.792

4 Conclusion and Future Work

In this paper, we presented our approach for the PAN 2018 author profiling shared task for predicting an author's gender using text-based and image-based features. We submitted a Logistic Regression classifier using word and character n-grams as text-based features and several automatically extracted image features. We found that only using text-based n-gram features gave us the best results for English and Spanish, whereas the combination of text-based and image-based features gave us the best results for Arabic. As additional text-based features we tested word embeddings, but results on the training data shows that these rather hurt our system's performance.

For this shared task we experimented with using images to predict an author's gender. We used an image feature extraction tool to classify detected faces in images on gender and emotion. We also tried to construct a feature that could indicate selfies, as

females tend to post more selfies than males. Our results showed that only using such image-based features are performing poorly with accuracy scores around 60% with a Logistic Regression classifier. Adding these features to a text-based n-gram model does not influence the score much. The images decreased the scores slightly on English and Spanish, but gave us a small improvement on Arabic on the PAN 2018 test dataset. Our submitted system only used image-based features extracted from detected faces, but data showed that not all images includes a face. Therefore, for future research we suggest a system that enlarges the set of image-based features.

References

1. Alrifai, K., Rebdawi, G., Ghneim, N.: Arabic tweeps gender and dialect prediction. Cappellato et al.[13]
2. Arriaga, O., Valdenegro-Toro, M., Plöger, P.: Real-time convolutional neural networks for emotion and gender classification. CoRR abs/1710.07557 (2017), http://arxiv.org/abs/1710.07557
3. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-gram: New groningen author-profiling model (2017), conference and Labs of the Evaluation Forum (CLEF 2017) : Information Access Evaluation meets Multilinguality, Multimodality, and Visualization ; Conference date: 11-09-2017 Through 14-09-2017
4. Ciobanu, A.M., Zampieri, M., Malmasi, S., Dinu, L.P.: Including dialects and language varieties in author profiling. arXiv preprint arXiv:1707.00621 (2017)
5. Dhir, A., Pallesen, S., Torsheim, T., Andreassen, C.S.: Do age and gender differences exist in selfie-related behaviours? Computers in Human Behavior 63, 549–555 (2016)
6. Kheng, G., Laporte, L., Granitzer, M.: Insa lyon and uni pasau's participation at pan@ clef'17: Author profiling task. Cappellato et al.[13]
7. Kodiyan, D., Hardegger, F., Neuhaus, S., Cieliebak, M.: Author profiling with bidirectional rnns using attention with grus: notebook for pan at clef 2017. In: CLEF 2017 Evaluation Labs and Workshop–Working Notes Papers, Dublin, Ireland, 11-14 September 2017 (2017)
8. Kuijper, M., Lenthe, M., Noord, R.: Ug18 at semeval-2018 task 1: Generating additional training data for predicting emotion intensity in spanish. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 279–285 (2018)
9. Markov, I., Gómez-Adorno, H., Sidorov, G.: Language-and subtask-dependent feature selection and classifier parameter tuning for author profiling. Working Notes Papers of the CLEF (2017)
10. Martinc, M., Skrjanec, I., Zupan, K., Pollak, S.: Pan 2017: Author profiling - gender and language variety prediction. In: CLEF (2017)
11. Noreen, E.W.: Computer-intensive methods for testing hypotheses. Wiley New York (1989)
12. Ogaltsov, A., Romanov, A.: Language variety and gender classification for author profiling in pan 2017. Cappellato et al.[13]
13. Oliveira, R.R., de Oliveira Neto, R.F.: Using character n-grams and style features for gender and language variety classification
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. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014), http://www.aclweb.org/anthology/D14-1162

16. Potthast, M., Pardo, F.M.R., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of pan'17 - author identification, author profiling, and author obfuscation. In: CLEF (2017)
17. 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)
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: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al. pp. 750–784 (2016)
19. Schaetti, N.: Unine at clef 2017: Tf-idf and deep-learning for author profiling. Cappellato et al.[13] (2017)
20. Sierra, S., Montes-y Gómez, M., Solorio, T., González, F.A.: Convolutional neural networks for author profiling. Working Notes Papers of the CLEF (2017)
21. Sorokowski, P., Sorokowska, A., Oleszkiewicz, A., Frackowiak, T., Huk, A., Pisanski, K.: Selfie posting behaviors are associated with narcissism among men. Personality and Individual Differences 85, 123–127 (2015)
22. op Vollenbroek, M.B., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., Nissim, M.: Gronup: Groningen user profiling (2016)