Combining textual and visual representations for multimodal author profiling

Notebook for PAN at CLEF 2018

Sebastian Sierra[1] and Fabio A. González[1]

Computing Systems and Industrial Engineering Dept., Universidad Nacional de Colombia
Bogotá, Colombia
{ssierral, fagonzalezo}@unal.edu.co

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. Author Profiling studies the common use of language inside those demographic groups. This work describes our proposed method for the PAN 2018 Author Profiling shared task. This year's task consisted of evaluating gender using multimodal information (text and images) which was extracted from Twitter users. We trained separate models for text, image and multimodal approaches. In multimodal approaches we explored early, late and hybrid approaches. We found experimentally that early approaches obtained the best performance. We obtained $0.80, 0.74$ and 0.81 of accuracy in the multimodal scenario for the test partition for English, Spanish and Arabic respectively.

1 Introduction

Nowadays, there is a large amount of information generated by users on various social networks. Facebook, Twitter, Instagram, among others generate a high amount of information in which users write their opinion about a topic, upload a photo about a relevant topic to them or simply record a video of what they are doing at that moment. Key applications can be derived from the generation of automatic analysis methods, which can handle properly the multimodal nature of social media information. Due to the increasing amount of social media information, several tasks for social media automatic analysis have acquired a greater importance. One of those tasks is Author Profiling (AP). AP can be seen as the study of the use of language in different demographic groups (profiles). For instance, gender-based profiles [3], age-based [22], native country-based [20], among others. Gender detection is one of the most popular subtask in Author Profiling [11, 22, 4, 8, 16, 13, 15, 18]. However, most of the work in AP has been devoted to the use of texts for categorizing correctly the profile of an author. Gender identification based on the images that an user posts in his/her social media is a task that has been gaining interest [6, 23, 30, 27]. Most of these works take the images of a social media user, extract the visual concepts, for instance, if a bag is present in the image and finally associate the presence of this concepts to the gender of the user (profile). Shigenaka et al. [23] interestingly propose a neural architecture which learns a proper representation for the images while associates it with the visual concepts which are extracted from the images.

1.1 Information fusion

Information fusion considers the problem of merging correctly two different representations of the same concept [5, 7].Atrey et al. [5] considers three levels of information fusion: feature level or early fusion, decision level or late fusion, and hybrid approaches. For this work, feature level consists of extracting text and visual representations and combining them into a single learning method. These combinations ignore the intrinsic correlation between modalities [17]. Decision level consists of combining the output decisions of previously learned classifiers for each modality. Hybrid approaches consist of methods that create a joint space for representing the different modalities of a concept, for instance for solving image captioning tasks [26, 28]. Multimodal approaches for Author Profiling have been considered by [1, 14, 25]. Álvarez-Carmona et al. [1] extend the PAN AP 2014 corpus by extracting a large set of tweets and images from the original users of this corpus. While their fusion strategy consists of an early fusion of text and image features. Taniguchi et al. [25] propose a hybrid fusion strategy, where visual concepts are extracted using a CNN, but each concept has a probability of being associated to a dimension of the profile (male or female). Text representation is extracted as the probability of a document to belong to a female user or a male user. At the very end, all the probabilities are concatenated and fed to a logistic regression classifier. It is worth to mention that these approaches are very recent and AP using multimodal strategies is becoming a very important topic in the scientific community.

1.2 PAN Author Profiling 2018 Shared Task

PAN-AP 2018 shared task consisted of classifying correctly the gender of an user of Twitter [19]. Two modalities were considered for representing an user: text and image. For each user, 100 tweets and 10 posted images were extracted. Also, users were selected from different languages: English, Spanish and Arabic. 1500 users were collected for the Arabic split, while 3000 users were collected separately for English and 3000 users for Spanish.

2 Methodology

In this section, we describe our submission to the PAN-AP 2018 shared task. Each subsection in this methodology describes the preprocessing, the feature extraction process and the learning algorithm used to classify an author as male or female.

2.1 Text representation

Twitter text representation strategies were explored at two levels. At the first level, several preprocessing strategies were considered: removing URLs, removing stopwords, lowercasing tweets, filtering retweets, usernames, hashtags and stripping of accents. Then, these main representations were explored:

– Bag-of-Words using unigrams (**WORD_UNI**): BoW representation is built using only unigrams with higher document frequency than 10 documents. Preprocessing steps for unigrams were URL removal, lowercasing, retweet filtering, usernames filtering, stopwords removal and accent striping.
– Bag-of-Words using bigrams (**WORD_BI**): Word bigrams with a higher document frequency than 10 documents were extracted from the training corpora. Preprocessing steps were the same as **WORD_UNI**, but stopwords were not removed.
– Bag-of-Words using character n-grams (**CHAR**): N-grams of characters were extracted using Scikit-Learn. 2-grams, 3-grams and 4-grams were used for representing stylistic features from the documents. Only preprocessing steps were: URL removal, hashtag and usernames filtering.
– Concatenation of all bag of words representations (**ALL_BOW**): **WORD_UNI**, **WORD_BI** and **CHAR** representations are concatenated.
– Average of fastText representations (**AVG_FAST**): Each word of each document was represented using a pretrained model of fastText [10]. Then, each author was represented by the feature-wise average of the word embeddings extracted from each fastText model. For each language a separate model was used.

2.2 Visual representation

Extracting visual features from a set of images is not an easy task. In the recent years, Convolutional Neural Networks (CNN) have gained a lot of attention by their competitive performance for several computer vision tasks. CNNs are built upon the idea of building high-level features using a compositional hierarchy of low-level features [12]. This means, first layers are expected to capture low-level patterns like edges, while higher layers are expected to learn domain specific features, which combine properly the low-level features. In image classification, the last layer of a CNN is commonly a SoftMax Layer with a size depending on the number of classes that it attempts to predict. Deep learning models like CNNs require a large amount of data for training, however they present two additional advantages: they can capture a pattern regardless of its location in the image, and the learned patterns can be transferred to solve a related image classification task. As described by Yosinski et al. [29], we can use the activation values of pre-trained CNNs for extracting features in images that belong to different classification domains. Since 2012, CNNs have been the state-of-the-art methods for image recognition tasks. In this work, we use two CNN architectures: VGG [24] and ResNet50 [9]. Both had a top performance in the ImageNet Large Scale Visual Recognition Challenge [21] during 2014 and 2015. Both VGG and ResNet50 are easy to use in Keras, so they were chosen as feature extractors:

– **ResNet50**: Receives a RGB image of size 224×224 and produces an output of 2048 non-negative values.
– **VGG16**: Receives a RGB image of size 224×224 and produces an output of 4096 non-negative values.

Before any image is fed to the network, they are scaled without losing their aspect ratio. Both networks produce non-negative features. For representing an author, all

his/her images were propagated through the network and the resulting feature vectors are averaged across each feature. This means, every author is represented by a vector of size 2048 or 4096, depending on the feature extractor.

2.3 Multimodal representation

As stated in Section 1, information fusion strategies can be categorized into: early fusion or feature fusion, late fusion or decision fusion, and hybrid approaches. In this work, we use one implementation of each category in order to assess the best strategy for fusing both modalities:

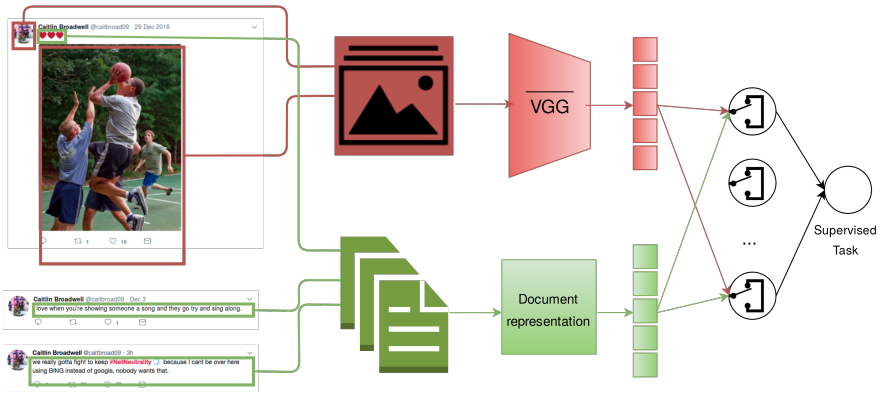

Figure 1. Methodology describing our approach to multimedia author profiling. Tweets from one author are gathered, then multimodal information is extracted from his/her tweets. Image representation is extracting using an average VGG16/ResNet50, while textual representation is built using bag representation. Finally, a multimodal representation is learned using GMUs.

- **Early Fusion (CONCAT)**: Best features from each modality (text and image) are stored, then each feature is scaled so it has zero mean and standard deviation of one. Finally a classifier is trained on top of these standardized features.
- **Late Fusion (VOTING)**: The best classifiers from each modality are stores using joblib library. This includes storing Scikit-Learn pipelines of transformation of data. Then, image and text are propagated through their respective classifier. Finally, the output probabilities are averaged and the max value is chosen as the predicted class.
- **Hybrid Fusion (GMU)**: For each language, a GMU [2] is trained using the best features per modality. Training split was divided again in training and development splits. The development split was used for validating the hyperparameters of the GMU. GMUs have the advantage of learning a a multimodal representation, while attempting to solve a supervised task (gender prediction). In Figure 1, we describe the methodology of our multimodal approach. Also, the best features in the textual modality had a large dimensionality, therefore a PCA was applied to retain th 99$ of variance.

3 Experiments and Results

For each language, the dataset was split into training and validation. 70% of the dataset was used for training and the remainder for validation. When the best performing features were extracted, a model was trained again using the complete dataset. The code for extracting the features was saved using joblib and was deployed in the TIRA evaluation system, where the evaluation on the test split was carried on. For the text modality, the following results were obtained using the proposed features.

Table 1. Results on validation for gender task using only the text modality

Text Features	**English**	**Spanish**	**Arabic**
WORD_UNI	0.79	**0.79**	0.55
WORD_BI	0.79	0.73	0.54
CHAR	**0.81**	0.76	**0.79**
AVG_FAST	0.70	0.68	
ALL_BOW	**0.81**	0.78	0.74

Table 2. Results on validation for gender task using only the image modality

Visual Features	**English**	**Spanish**	**Arabic**
ResNet50	**0.78**	**0.77**	**0.7**
VGG16	0.75	0.7	0.7

Table 3. Results on validation for gender task using multimodal approaches

Multimodal Approaches	**English**	**Spanish**	**Arabic**
CONCAT	**0.82**	**0.80**	**0.79**
VOTING	0.80	0.75	0.78
GMU	0.81	0.77	**0.79**

4 Discussion and Conclusion

Character n-grams worked very well for identifying gender on English and Arabic, as can be seen on Table 1. While for Spanish, the word unigrams worked better. In the image modality, **ResNet50** outperformed **VGG16** as a feature extractor as can be seen on Table 2. However the strategy for combining the extracted features of the images of one author consisted only of a feature-wise average.

One of the main motivations of the work, was to learn a multimodal representation using GMUs. Strong regularization using dropout and batch normalization per modality was used for training GMUs. Although for the case of English, only 2100 samples were fed to the GMU. While for Arabic, the number of samples decreased to 1050. Although we used strong regularization, our model overfitted quickly. In Table 3 we showed that early fusion approaches obtained the best results for the multimodal task.

Future work involves improving the way that image representations are extracted and combined. Taniguchi et al. [25] provide a very good strategy for extracting visual information from the posted images. Also, we are interested in applying successfully GMUs to the Author Profiling task.

Bibliography

[1] Álvarez-Carmona, M.A., Pellegrin, L., Montes-Y-Gómez, M., Sánchez-Vega, F., Escalante, H.J., López-Monroy, A.P., Villaseñor-Pineda, L., Villatoro-Tello, E.: A visual approach for age and gender identification on Twitter. Journal of Intelligent & Fuzzy Systems 34(3133-3145), 1–5 (2018)

[2] Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated Multimodal Units for Information Fusion. In: International Conference on Learning Representations ICLR 2017 - Workshop. Toulon, France (feb 2017), http://arxiv.org/abs/1702.01992

[3] Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, Genre, and Writing Style in Formal Written Texts. Text - Interdisciplinary Journal for the Study of Discourse 23(3) (2003)

[4] Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically Profiling the Author of an Anonymous Text. Communications of the ACM 52(2), 119–123 (2009), http://doi.acm.org/10.1145/1461928.1461959

[5] Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia Systems 16(6), 345–379 (apr 2010), http://dx.doi.org/10.1007/s00530-010-0182-0 http://link.springer.com/10.1007/s00530-010-0182-0

[6] Azam, S., Gavrilova, M.: Gender prediction using individual perceptual image aesthetics (2016), https://otik.zcu.cz/handle/11025/21646

[7] Bhatt, C., Kankanhalli, M.: Multimedia data mining: state of the art and challenges. Multimedia Tools and Applications 51(1), 35–76 (2011), http://dx.doi.org/10.1007/s11042-010-0645-5

[8] Burguer, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on Twitter. In: Proceedings of the conference on empirical methods in natural language processing (2011), https://dl.acm.org/citation.cfm?id=2145568

[9] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[10] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification (2016), http://arxiv.org/abs/1607.01759

[11] Koppel, M., Argamon, S., Shimoni, A.R.: Automatically Categorizing Written Texts by Author Gender. Literary and Linguistic Computing 17(4), 401–412 (nov 2002), http://dx.doi.org/10.1093/llc/17.4.401

[12] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436 (2015), https://www.nature.com/articles/nature14539

[13] 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 (nov 2015), http://www.sciencedirect.com/science/article/pii/S0950705115002427

[14] Merler, M., Liangliang Cao, Smith, J.R.: You are what you tweet...pic! gender prediction based on semantic analysis of social media images. In: 2015 IEEE

International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (jun 2015), http://ieeexplore.ieee.org/document/7177499/
[15] Ortega-Mendoza, R.M., López-Monroy, A.P., Franco-Arcega, A., Montes-y Gómez, M.: Emphasizing personal information for Author Profiling: New approaches for term selection and weighting. Knowledge-Based Systems 145, 169–181 (apr 2018), https://www.sciencedirect.com/science/article/pii/S0950705118300224
[16] Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting Age and Gender in Online Social Networks. In: Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents. pp. 37–44. SMUC '11, ACM, New York, NY, USA (2011), https://dl.acm.org/citation.cfm?id=2065035 http://doi.acm.org/10.1145/2065023.2065035
[17] Pei, D., Liu, H., Liu, Y., Sun, F.: Unsupervised multimodal feature learning for semantic image segmentation. In: The 2013 International Joint Conference on Neural Networks (IJCNN). pp. 1–6. IEEE (aug 2013), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748
[18] Rangel, F., Rosso, P.: On the impact of emotions on author profiling. Information Processing & Management 52(1), 73–92 (jan 2016), http://www.sciencedirect.com/science/article/pii/S0306457315000783
[19] 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)
[20] Rangel Pardo, F.M., 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, vol. 1866. CLEF and CEUR-WS.org (2017), http://ceur-ws.org/Vol-1866/
[21] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115(3), 211–252 (dec 2015), http://link.springer.com/10.1007/s11263-015-0816-y
[22] Schler, J., Koppel, M., Argamon, S., Pennebaker, J.J.: Effects of Age and Gender on Blogging. In: AAAI spring symposium: Computational approaches to analyzing weblogs. pp. 199–205 (2006), http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-039.pdf
[23] Shigenaka, R., Tsuboshita, Y., Kato, N.: Content-Aware Multi-task Neural Networks for User Gender Inference Based on Social Media Images. In: 2016 IEEE International Symposium on Multimedia (ISM). pp. 169–172. IEEE (dec 2016), http://ieeexplore.ieee.org/document/7823607/
[24] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
[25] Taniguchi, T., Sakaki, S., Shigenaka, R., Tsuboshita, Y., Ohkuma, T.: A Weighted Combination of Text and Image Classifiers for User Gender Inference. In:

Proceedings of the 2015 Workshop on Vision and Language (VL'15). pp. 87–93. Lisbon, Portugal (2015), http://www.anthology.aclweb.org/W/W15/W15-2814.pdf
[26] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and Tell: A Neural Image Caption Generator. CoRR (nov 2014), http://arxiv.org/abs/1411.4555
[27] Xiaojun Ma, Tsuboshita, Y., Kato, N.: Gender estimation for SNS user profiling using automatic image annotation. In: 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). pp. 1–6. IEEE (jul 2014), http://ieeexplore.ieee.org/document/6890569/
[28] Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. arXiv preprint arXiv:1502.03044 (2015)
[29] Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, vol. 27, pp. 3320–3328. Curran Associates, Inc. (2014), http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.pdf
[30] You, Q., Bhatia, S., Sun, T., Luo, J.: The Eyes of the Beholder: Gender Prediction Using Images Posted in Online Social Networks. In: 2014 IEEE International Conference on Data Mining Workshop. pp. 1026–1030. IEEE (dec 2014), http://ieeexplore.ieee.org/document/7022709/