A Straightforward Multimodal Approach for Author Profiling

Notebook for PAN at CLEF 2018

Mario Ezra Aragón[1] and A. Pastor López-Monroy[2]

[1] Department of Computer Engineering,
Universidad Autónoma de Chihuahua; Chihuahua, Chih., México, 31100.
[2] Department of Computer Science,
University of Houston; Houston TX, USA 77004
p235736@uach.mx, alopezmonroy@uh.edu

Abstract In this paper we evaluate different strategies from the literature for text and image classification at PAN 2018. The main objective of this shared task is the identification of the gender of different users by using tweets and images posted. We evaluate four popular strategies for the text representation: 1) Bag of Terms (BoT), 2) Second Order Attributes (SOA) representation, 3) Convolutional Neural Network (CNN) models and 4) an Ensemble of n-grams at word and character level. For the image representation we used a Convolutional Neural Network (CNN) based on [6]. We observed that the n-grams Ensemble presented the highest performance. For our participation we chose the Ensemble and perform an early fusion with the image representation to create a multimodal representation.

Keywords: Author Profiling · Bag of Words · CNN · Text Classification · Text Mining

1 Introduction

Author Profiling (AP) is a common and well know task in Natural Language Processing (NLP), that consists in extracting all the possible information from an author's document [11]. The AP could help in different problems such as the detection of a person of interest, security, prevention, political opinion, business intelligence, etc. The PAN 2018 shared task has the objective of tackling this problem using machine learning and NLP techniques. The main objective is to identify the gender's user with the novelty of considering posted tweets and images as new information. The shared task has three different languages: English, Spanish and Arabic.

In this work we separately evaluate the AP in three modalities: Textual, Visual and Textual-Visual. For the textual modality, we mainly evaluate what should be the *de facto* baseline: a huge ensemble of n-gram histograms at word and character level. Then, we compare this approach with three strategies: Bag-of-Terms [3], Second Order Attributes [4] and CNNs [5]. The core idea behind our evaluation is to determine which approach captures better the thematic content, which according to the literature has been the cornerstone to effectively profile users [11,13,14]. Regarding to the visual modality, we only evaluate one very simple, yet effective, CNN based method.

For this visual approach, we extracted the category layer from the VGG16 [6] and use it as features with a SVM on the top. Then, a set of images belonging to the same users are averaged. Intuitively this approach exploits the posting behavior of users [12], where the idea is to capture the visual content that is being posted by users. The intuition is that such visual content is significantly different between males and females, thus highly discriminative. This is somewhat analogous to observe the thematic content when classifying documents. Finally for the Textual-Visual modality, we bring together the textual-visual thematic into a single approach. For this we performed an early fusion to combine our multiple histograms of n-grams and the features extracted from the VGG16; this is precisely our submitted approach to PAN18.

The remainder of this paper is at follows: Section 2 presents some of the AP related work. In Section 3 we described the different strategies that we evaluated for the text representation. Section 4 describes our approach for the image representation. In Section 5 is described the multimodal representation and how was created. Section 6 and 7 describes the Experimental Settings and a description of the Experimental Results. Finally Section 8 include our conclusions.

2 Related Work

In this section we present a review of AP related work that have been proposed to handle this task. There are different methods: from a simple representations like removing stopwords and creating a Bag of Terms [17] to a more complex representations using traditional word embeddings [20] or embeddings exploiting the morphology and semantics of the words [21]. In [19] the authors proposed a simple method of classification based on the similarity between the objects; they consider different terms used in the texts that corresponds to a user's tweets. Other approach is to extract groups of terms that are presented in the tweets [13,22], where the authors also used extra information like emojis, document sentiment, POS tags, etc. In these approaches the authors found that including the extra information like emojis or POS tags do not improve the performance.

Another popular approach is to address the problem as a profile based problem [18,4], where they create targets of profiles and groups of subprofiles for each user's tweets. In [16] authors built a system where they used a combination of different classifiers, with the objective of identified the behavior of different users. There are also some approaches that handle this task using relative new approaches like deep learning. For instance in [21] the authors generate embeddings representations that are classified using deep averaging networks. This model receives as input the word embeddings and the first layer average those embeddings, the next hidden layers transform the computed average. In [23] the authors used a deep learning model based on CNNs using a matrix of 2-grams of letters with punctuation marks as features. These deep learning approaches got an accuracy above the average results for the task.

When we talk about visual and multimodal for AP, these approaches have been less studied in comparison with text approaches. For the visual modality approach, the authors had focused their research on gender recognition task [24,25,26], where some general statistics have been considered using the images as features. For the multimodal

approach, is a type of strategy that has just recently been explored [27]. In [28] authors used a image and text weighted strategy for gender classification. Their idea consist using a CNN for determining a score related to a user's image, and they combined this information with textual features and average the score.
In this work, inspired in the related work we attempt to evaluate different approaches (see Section 3) and select the best one for the test data presented for the task. We proposed to bring an early fusion from the textual and visual features for our approach.

3 Textual Modality Strategies

In this section we described the different strategies that we select from the literature for the textual representation. All these representations have resolved and got remarkable performance in different NLP classification tasks.

3.1 Bag of Terms (BoT)

The bag of terms is the most simple and well know strategy for text representation where the text is described by the occurrence of words within a documents, i)the first step is the creation of a vocabulary form training data and then ii) the presence of the words are measure by its frequency [3]. This representation is an histogram thus it ignores the structure of the words, accounting only the occurrence of the words in the document and not the position or order in it.

3.2 Second Order Attributes (SOA)

In this representation the document vectors are build in a space of profiles. Where each value in the vector represents the relationship that exist between each document with each target profile and subprofiles [4]. This representation has the objective of dividing the profiles using a clustering algorithm to create several subprofiles. First is needed to capture the relation of each term with the profiles. Then compute the term vector in a profile's space, it creates a term vectors of the terms that are contained in the document. Lastly they are weighted by the relative frequency of the term contained in the document.

3.3 CNN Models

For this strategy we used CNN models that are based on [5], we used three different training techniques for the models:

- CNN-Rand: This model tested is where we randomly initialized all weights and then are modified during the training phase..
- CNN-Static: Where this model uses word embedding vectors to initialize the embedding layer. During the training the weights of the embedding are kept fixed so they are not modified.

– CNN-NonStatic: This model is similar as the previous one, but we allowed to change the embedding weights during training.

We used filter windows of size 3,4 and 5 with 100 feature maps for each one, a dropout rate of 0.5, and stochastic gradients descent for training over shuffled mini-batches. We used pre-trained word vectors with a dimensionality 300, word vectors were obtained using word2vec [10] for English and FastText [9] for Spanish and Arabic.

3.4 N-Grams Subspaces for Author Profiling

The first step for our approach is the creation of the representation from the text. We proposed a method that has two stages for the creation: i) extract n-grams of size one to four at word level and size two to five at character level, ii) then we select the best n-grams using chi^2 distribution applied to each group and then concatenate the best selected n-grams from each group, as shows in Figure 1 we can see the overall process of extraction and creation of the n-grams. In the following lines we explain the main two stages.

Extract n-grams The first step for our approach is to create the group of n-grams [2] of size one to four for the word level and two to five for the character level. To extract the n-grams we have three steps i) first we represent the documents using the occurrences of the group of words in the document, ii) then each group of terms vectors are normalize and iii) smoothing the weights of each group by the inverse document frequency adding one to document frequencies, preventing zero divisions as if every terms was seen in other document.

CHI2 distribution The second stage of this approach is the selection of the best features of each group of n-grams, we used the chi^2 distribution X_k^2 [1] for this selection. When using this function we select the features that are the most likely to be relevant for the detection of the gender.

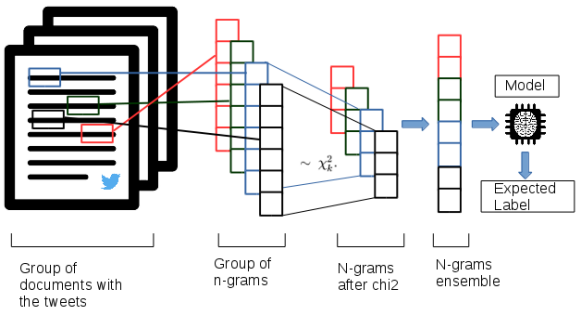

Figure 1. N-gram Ensemble diagram creation

4 Visual Modality: CNN for features extraction

The second step of our approach is the feature extraction of the images. Each user has 10 images that they post in the social media. We use a well know state-of-the-art model in [6] with pre-trained weights on ImageNet. We used the last layer (the class layer with the 1000 classes) of the pre-trained model as the features of the feed image. Then create a mean vector formed from the features of the 10 images. These mean vectors are used for training the image model. The idea behind this approach is to capture a similar distribution of images that users post, and achieve a discrimination between them. As the model is designed for visual object recognition (this includes objects and scenes), we expect similar values for users that post similar images. The pre-trained model that we used was the VGG16 with 1000 classes, VGG16 is a CNN model and refers to the 16 weight layers of the model. Figure 2 shows the extraction of the features from the images for training the image model.

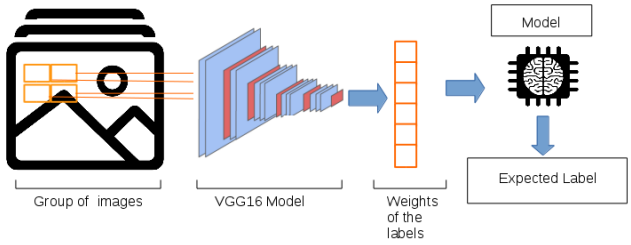

Figure 2. Image feature extraction diagram

5 Early Fusion: Textual and Visual Representations

The last step for the shared task is the classification of the users using both text and images. Our approach consist of an early fusion of both Textual and Visual representations concatenating previous vectors and then we pass the new representation to the classifier, we used a Support Vector Machine (SVM) for the training and classification. Our hypothesis is that combining both features the results should improve by giving more information about the users, than only using one kind of feature. Figure 3 describes this feature extraction from the text and the images for the concatenation.

6 Experimental Settings

The objective for this task is to determinate the gender of a user using a set of different tweets and images that the user posted. We evaluate the task in a separated way: i) only using the text we extract the group of n-grams and trained the SVM classifier for the

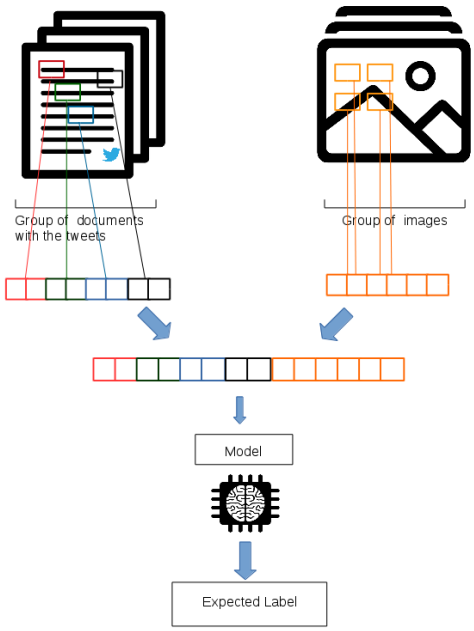

Figure 3. Diagram of Early Fusion with Textual and Visual Representations

prediction, ii) only using the images we used the VGG16 to extract the features of the images and calculate the mean vector of it and then trained other SVM classifier, and iii) we used both the features of the text and the images then concatenate and trained a third SVM classifier. The shared task have 3 different languages to test the models, and we trained one model for each language with this we have 9 different predictions for the test dataset. In [15] it presents an overview that describe in detail the tasks, data and evaluation.

7 Experimental Results

To test the models before the TIRA platform we first separated the training dataset in 70% for training and 30% for test and extract the text and images features. For this task we measure the Accuracy over the predictions. For the test dataset we trained our models and then predict the gender using all the users in the training dataset. Table 1 shows the detailed classification results obtained with the text, images and both strategies for the three languages. In these results we could appreciate that the n-gram Ensemble performs better than the others strategies for text representation. SOA and CNN did not perform better in these tasks than the n-grams Ensemble this could be due the base term (words) used in those representations. Therefore this presents an opportunity to integrate the same idea as the n-grams and look for a better performance. The image representation alone did not perform better than only using the text, but when we com-

bine both representations it increases the results.

Table 1. Detailed classification accuracy

Language	Training Data								Test Data		
	Text						Image	Text and Image	Text	Image	Text and Image
	BoT	SOA	CNN-Rand	CNN-Static	CNN-NonStatic	N-Gram Ensemble	VGG16	N-Gram + VGG16	N-Gram	VGG16	NG+VGG16
English	0.7778	0.7717	0.5727	0.6182	0.6010	0.8495	0.6848	**0.8515**	0.7963	0.6921	0.8016
Spanish	0.7667	0.7485	0.5768	0.5879	0.6414	0.8414	0.6879	**0.8465**	0.7686	0.6668	0.7723
Arabic	0.5798	0.5980	0.5354	0.5394	0.5354	0.8343	0.6949	**0.8444**	0.6480	0.6800	0.6670

In order to study the remarkable performance of the n-grams, we extract the best 10 n-grams for the words group from the English and Spanish training corpus that were obtained using the chi^2 distribution and then cherry pick the best 5, Table 2 shows these group of words. In this table we can appreciate the selection of words that people prefer to use when they tweet about something of their interest.

Table 2. Best n-grams at word level for English and Spanish users

English				Spanish			
1 word	2 words	3 words	4 words	1 word	2 words	3 words	4 words
'cute'	'More for'	'have the best'	'have the best day'	'amiga'	'mi novio'	'el gol de'	'de Trump https co'
'girls'	'my bed'	'in the league'	'liked YouTube video from'	'equipo'	'te amo'	'en mi corazón'	'EE UU https co'
'league'	'my mum'	'so excited to'	'new photo to Facebook'	'gol'	'gol de'	'más grande de'	'en EE UU https'
'lovely'	'my wife'	'happy birthday mate'	'photo to Facebook https'	'jugador'	'un equipo'	'porque no me'	'la vida https co'
'mum'	'the league'	'have lovely day'	'posted new photo to'	'partido'	'mi corazón'	'que mi mamá'	'que si https co'

To analyze the performance of the image model, we select some images and get the probabilities from our model of been post by male or female. Figure 4 shows the probabilities from some pictures from the three languages. For the English users, we can appreciate that sport's images related are more common for males and landscapes or cats are more common for females. We also present images from the Spanish users where for male is more common to post about sports and video games and for the females their prefer pictures from artist and landscapes. Last part of the figure shows images from the Arabic user where we can appreciate that males have a high probability of posting something related to sports too (even greater than English and Spanish) and for females is common to post more colorful pictures. But in general there are a lot of neutral pictures about politics, social events or comic images that are harder to classify.

8 Conclusions

In this notebook we presented an approach in order to determine the gender of a user using the tweets and images they post. For the text part we used four different strategies, where the n-gram Ensemble gets the best overall performance. We used a n-gram

Figure 4. Probabilities of the images of been post by male or female.

representation and analyze it to see the different group of words that have most discriminative value to give weight to the classes, in this representation we could see that the model capture important words that the users post when they tweet about something. For the image part we used a VGG16 model to extract features from the images and capture the kind of image that people usually post. The images alone did not get the expected results, due the similarity of the image topics about politics or social events. Then for the final step we concatenate the features from the text and images to see if the model could gain extra information for the classification. With these experiments we obtained evidence that only text information gives better results than only using the images, but the features combined improves the results in the training and test sets proving our hypothesis.

References

1. Walck, C.: Hand-book on Statistical Distributions for experimentalists. Internal Report SUF–PFY/96–01, Stockholm (2007)
2. Jurafsky, D., Martin, J.: Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Third Edition draft. Chapter 4, (2014)
3. Goldberg, Y.: Neural Network Methods in Natural Language Processing (Synthesis Lectures on Human Language Technologies). Graeme Hirst, (2017)
4. Lopez-Monroy, A.P., Montes-Y-Gomez, M., Escalante, H.J., Villasenor-Pineda, L., Villatoro-Tello, E.: Inaoe's participation at pan'13: Author profiling task. In: Notebook Papers of CLEF 2013 LABs and Workshops, Valencia, Spain, (September 2013)
5. Kim, Y: Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), (2014)
6. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, (2014)
7. Yin, W., Kann, K., Yu, M., Schütze, H.: Comparative Study of CNN and RNN for Natural Language Processing. CoRR, (2017)
8. Moreno-Lopez, M., Kalita, J.: Deep Learning applied to NLP. CoRR, (2017)
9. Joulin, A., Grave, E., Bojanowski, P., Mikolovn, T.: Bag of Tricks for Efficient Text Classification, CoRR (2016)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space, CoRR (2013)
11. Rangel, F., Rosso, P.: Use of language and author profiling: Identification of gender and age. Natural Language Processing and Cognitive Science, page 177 (2013)
12. Álvarez-Carmona, M., 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. IOS Press (2017)
13. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-GrAM: New Groningen Author-profiling Model. Notebook for PAN at CLEF (2017)
14. 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 (2017)
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. Agrawal, M., Goncalves, T.: Age and Gender Identification using Stacking for Classification. Notebook for PAN at CLEF 2016 (2016)
17. Bakkar-Deyab, R., Duarte, J., Gonçalves, T.: Author Profiling Using Support Vector Machines. Notebook for PAN at CLEF 2016 (2016)
18. Bougiatiotis, K., Krithara, A.: Author Profiling using Complementary Second Order Attributes and Stylometric Features. Notebook for PAN at CLEF 2016 (2016)
19. Adame-Arcia, Y., Castro-Castro, D., Ortega-Bueno, R., Muñoz, R.,: Author Profiling, instance-based Similarity Classification. Notebook for PAN at CLEF 2017 (2017)
20. Akhtyamova, L., Cardiff, J., Ignatov, A.: Twitter Author Profiling Using Word Embeddings and Logistic Regression. Notebook for PAN at CLEF 2017 (2017)
21. Franco-Salvador, M., Plotnikova, N., Pawar, N., Benajiba, Y.: Subword-based Deep Averaging Networks for Author Profiling in Social Media. Notebook for PAN at CLEF 2017 (2017)
22. Martinc, M., Škrjanec, I., Zupan, K., Pollak, S.: PAN 2017: Author Profiling - Gender and Language Variety Prediction. Notebook for PAN at CLEF 2017 (2017)
23. Schaetti, N.: UniNE at CLEF 2017: TF-IDF and Deep-Learning for Author Profiling. Notebook for PAN at CLEF 2017 (2017)
24. Azam, S., Gavrilova, M.: Gender prediction using individual perceptual image aesthetics. Journal of WSCG, 24(2):53–62 (2016)
25. Ma, X., 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), pages 1–6, July (2014)
26. Hum, N. J., Chamberlin, P. E., Hambright, B. L., Portwood, A. C., Schat, A. C., Bevan, J. L.: A picture is worth a thousand words: A content analysis of Facebook profile photographs. Computers in Human Behavior, 27(5):1828 – 1833, (2011)
27. Merler, M., Cao, L., 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), pages 1–6, June (2015)
28. Taniguchi, T., Sakaki, S., Shigenaka, R., Tsuboshita, Y., Ohkuma, T.: A Weighted Combination of Text and Image Classifiers for User Gender Inference, pages 87–93. Association for Computational Linguistics (2015)