ISSN 2007-9737

Author Profiling in Social Media with Multimodal Information

Miguel Á. Álvarez Carmona[1,2], Esaú Villatoro Tello[4], Manuel Montes y Gómez[3], Luis Vilaseñor Pineda[3]

[1] Consejo Nacional de Ciencia y Tecnología, Mexico

[2] Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad de Transferencia Tecnológica, Mexico

[3] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Laboratorio de Tecnologías del Lenguaje, Mexico

[4] Universidad Autónoma Metropolitana Unidad Cuajimalpa, Departamento de Tecnologías de la Información, Mexico

malvarez@cicese.mx

Abstract. This paper summarizes the thesis: "Author Profiling in Social Media with Multimodal Information." Our solution uses a multimodal approach to extracting information from written messages and images shared by users. Previous work has shown the existence of useful information for this task in these modalities; however, our proposal goes further, demonstrating the complementarity of the modalities when merging these two sources of information. To do this, we propose to transform images to texts, and with them, to have the same framework of representation for both kinds of information, which allow to achieve their fusion. Our work explores different methods for extracting information either from the text and the images. To represent the extracted information, different distributional term representations approaches were explored in order to identify the topics addressed by the user. For this purpose, an evaluation framework was proposed in order to identify the most appropriate method for this task. The results show that the textual descriptions of the images contain useful information for the author profiling task, and that the fusion of textual information with information extracted from the images increases the accuracy of this task.

Keywords. Author profiling, multimodal information, natural language processing, text classification.

1 Introduction

The author profiling task (AP) is to extract *demographic aspects* of a person from their texts. For example gender, age, location, occupation, socio-economic level or native language [21, 41]. Efforts have also been made to determine other aspects such as the level of well-being [42], personality traits such as extraversion or neuroticism [40, 39] as well as political ideology [19], an affinity for some products [7], among others [13].

In the AP context, it can be seen that most of the recent works, in the field of social networks, have focused mainly on the definition of thematic attributes and style-metrics appropriate for this task.

However, there is a sign of progress towards the description of multimodal representations that, for example, integrate different types of information. Due to the nature of social networks, the images shared by users or their social environment are also incorporated. This thesis work particularly considered the author profiling in social networks with multimodal information [32].

1.1 Problem

Most of the works that have tried to solve the task of AP are based solely on the textual information that users share on social networks. Utilize only text generates that much of the information available by the nature of social networks is not exploited. Most approaches do not take advantage of images, videos, contact lists, activity schedules, or other information. For this reason, it is not known which of these different information modalities is more valuable for the AP task. This is why it is essential to analyze how multimodal information impacts the AP task.

Another aspect to highlight is that the works in AP have given evidence of the importance of the content of the texts. Nevertheless, the most common approach that has been used is the Bag of Words (BoW). The problem with this approach when working on social networks is the lack of information because regular short texts are analyzed. Besides that, the texts are not formal, which causes that there are words out-of-the-dictionary and spelling mistakes.

A set of approaches that have not been deepened enough to represent the content of the texts and, that can be useful for the AP task are the **distributional term representations** (**DTRs**). The basic intuition behind the DTR's is called the distributional hypothesis [44], which states that terms with similar distributional patterns tend to have the same meaning [23, 25]. This distributional hypothesis could capture the content of the users' text in a better way than the traditional content approaches used for AP. In this thesis work, we compared these representations experimentally to know their impact on the AP task.

On the other hand, few works have taken advantage of the information extracted from the images shared by users, this even though various works in psychology have concluded that the photos that are shared on social networks can tell a lot of the people [15, 10, 50, 17]. Some works have applied the color histogram of the images to determine the gender of the users, but no studies have been done for other traits of the authors. Other works have converted the images to texts with automatic labelers of images, through **automatic images annotation** techniques, that assign a list of labels from a previously established set, and from there, infer the user's profile.

These approaches are commonly supervised and with a **closed vocabulary**. This means that the labelers select from a limit list of labels the elements in each image. The problem is that a limited vocabulary could be insufficient to represent the interest of the profiles in a collection. In this thesis work, we proposed to apply an approach based on **open vocabulary** to the AP task, under the idea that it describes in a better way the social media profiles. The automatic annotation of images based on open vocabulary approaches does not select the labels set from a limit list, but they select the vocabulary from an extensive collection, usually extracted from Internet pages. With this idea, we could represent each image in the collection as a text, and we were able to apply text processing approaches to classify the profile of each user.

1.2 Research Questions

Throughout this thesis, we intend to answer the following research questions:

1. What kind of information could be captured by distributional-based methods, and how effective are they for representing user's information when facing the problem of author profiling?
2. How to extract information from the images shared by users through an open vocabulary approach, and how to use them to determine their profile?
3. How to jointly take advantage of the information obtained from texts and images for solving the author profiling task?

1.3 Contributions

The main contributions derived from this work are:

1. A novel corpus including information about Mexican twitter accounts with text and image information. Also, the extension of the well knows PAN@14 corpus. This collection had only text information. For this study, we include the image information for this collection.
2. A comparison among different distributional based methods for the AP task. For this study, we apply DOR, TCOR, word2vec, and SSR.
3. A multimodal method for the AP task taking advantage of textual and image information.
4. The evidence that it is possible to classify profiles from different countries and language trough the different images shared on the networks.

In the following sections, we describe each of the main contributions of this work.

2 Corpora

We presented two new corpora that have been designed for the Author Profiling task evaluation with text and image information.

First, we presented an extension of the well known PAN 14 Twitter corpus [38], aiming to use a well-known corpus enriching it with image information.

Also, the thesis presented a Mexican Twitter corpus for the AP task. The specific application of this corpus is in the analysis of several traits of Mexican Twitter users by text and image information. The data contains for each account the activity schedule on Twitter, its tweets, and its images. This corpus is labeled for gender, the place where he/she lives, and occupation. The annotation of the data was been accomplished manually.

The rest of this chapter is organized as follows. Section 2.1 describes the PAN 14 corpus for the text experiments. Section 2.2 shows the description of the images extension for the PAN 14 Twitter corpus.

Table 1. Distribution of the gender and age classes across the different social media domains

Classes	Genres			
	Blogs	*Reviews*	*Social-media*	*Twitter*
Female	73	2080	3873	153
Male	74	2080	3873	153
Total:	147	4160	7746	306
18-24	6	360	1550	20
25-34	60	1000	2098	88
35-49	54	1000	2246	130
50-64	23	1000	1838	60
65+	4	800	14	8
Total:	147	4160	7746	306

Finally, Section 2.3 describe the new Mexican Twitter corpus for the author profiling task.

2.1 Pan 14 Corpus

For our experiments, we employed the English dataset from the PAN 14 AP track. This corpus was specially built for studying AP in social media. It is labeled by gender (i.e., female and male), and five non-overlapping age categories (18-24, 25-34, 35-49, 50-64, 65+). Although all documents are from social media domains, four distinct genres were provided: blogs, social media, hotel reviews, and Twitter posts. A more detailed description of how these datasets were collected can be found in [38]. Table 1 provides some basic statistics regarding the distribution of profiles across the different domains (i.e., genres). It can be noticed that gender classes are balanced, whereas, for the age classification task, the classes are highly unbalanced. Notably, there are very few instances for the 65+ category.

2.2 Extended PAN 14 Corpus

Images shared by social media users tend to be strongly correlated with their thematic interests as well as to their style preferences. Motivated by these facts, we tackled the task of assembling a corpus considering text and images from Twitter users. Mainly, we extended the PAN-2014 [38] dataset by obtaining images from the already existing Twitter users.

Table 2. Statistics of images shared by each age category

Ages	Profiles	Average images (α)	Average tweets (α)
18-24	17	246.45 ($\pm$80.34)	706.18($\pm$361.76)
25-34	78	286.42 ($\pm$202.65)	796.01($\pm$291.18)
35-49	123	301.74 ($\pm$253.83)	640.41($\pm$362.28)
50-64	54	334.19 ($\pm$238.24)	527.68($\pm$354.24)
65+	7	441.65 ($\pm$102.52)	651.85($\pm$432.28)

Table 3. Statistics of images shared by each gender category

Ages	Profiles	Average images (α)	Average tweets (α)
Female	140	162.21 ($\pm$294.13)	543.53($\pm$395.93)
Male	139	141.76 ($\pm$274.98)	784.88($\pm$265.86)

The PAN-2014 dataset includes tweets (only textual information) from English users. Based on this dataset, we obtained more than 42,000 images, corresponding to a subset of 279 profiles in English[1]. The images associated with all of the users were downloaded to existing user profiles, resulting in a new multimodal Twitter corpus for the AP task. Each profile has an average of 304 images.

Tables 2 and 3 present additional statistics on the values that both variables, gender and age can take, respectively. On the one hand, Table 2 divides profiles by age ranges, i.e., 18-24, 25-34, 35-49, 50-64 and 65+. It shows a great level of imbalance, being the 35-49 class, the one having the greatest number of users.

Nonetheless, the users from the 65+ range are the ones with the greatest number of posted images as well as the lower standard deviation values. It is also important to notice that the users belonging to the 50-64 range share in average a lot of images, but show a large standard deviation, indicating the presence of some users with too many and very few images.

On the other hand, Table 3 reports some statistics for each gender profile. It is observed a balanced number of male and females users in both corpora as well as a similar number of shared images.

[1]Note that the PAN-2014 corpus includes more profiles; however, for some Twitter users, it was impossible to download their associated images.

Fig. 1. Regional division for Mexico. Source: `http://www.conafor.gob.mx/`

2.3 Mex-A3T-500 Corpus

To study the characteristics of the different Mexican Twitter profiles, we built a Mexican corpus for author profiling named Mex-A3T-500[2]. Each of the Twitter users was labeled with gender, occupation, and place of residence information. For the occupation label, we considered the following eight classes: arts, student, social, sciences, sports, administrative, health, and others. For the place of residence trait, we considered the following six classes: north (norte), northwest (noroeste), northeast (noreste), center (centro), west (occidente), and southeast (sureste). Figure 1 shows the division in Mexico's map.

2.3.1 Construction of the Corpus

Two human annotators, working three months each, were needed for building this corpus. They applied the following methodology: (i) to find a set of Twitter accounts corresponding to famous persons and/or organizations from each region of interest. These accounts usually were from local civil authorities, known restaurants, and universities; (ii) to search for followers of the initial accounts, assuming that most of them belong to the same region with the initial accounts; (iii) to select only those followers that explicitly mention, in Twitter or another social network (as

[2]This is a subset of the corpus used for the MEX-A3T forum for the 2018 and 2019 editions [2, 6]. https://sites.google.com/view/mex-a3t/.

Table 4. Example of tweets mentioning information related to the place of residence and/or occupation of users

Trait detected	Original text	Translation
Residence	La pura carnita asada en **Monterrey**	Roast beef in **Monterrey**
Residence	Nunca me canso de pasear en el zócalo de **Puebla**	I never get tired of walking in the **Puebla** Zocalo
Occupation	Porque los **arquitectos** nunca descansamos	Because we, the **architects** never rest
Occupation	**Programando** en el trabajo ando	**Programming** at work

Table 5. Mexican author profiling corpus: distribution of the gender trait

Class	Profiles	Average images (α)	Average tweets (α)
Female	250	715.46 ($\pm$722.89)	1225.00($\pm$868.17)
Male	250	480.90($\pm$459.36)	1500.01($\pm$946.66)

Facebook and Instagram) their place of residence and occupation. Table 4 shows some examples of tweets where users reveal information from their place of residence and occupation.

2.3.2 Statistics

The corpus consists of 500 profiles from Mexican Twitter users. Each profile is labeled with information about the gender, occupation, and place of residence of the user. Tables 5, 6 and 7 present additional statistics on the distribution of user accounts on gender, occupation and location.

Table 6 divides profiles into the different Mexican regions on the corpus, i.e., north, northeast, northwest, center, west, and southeast. Also, it shows an important level of imbalance, being the center class, the one having the greatest number of users, while the north is the class with the lowest.

On the other hand, Table 7 divides profiles on the eight different occupations on the corpus. It is possible to see that the majority class is the central region, whereas the classes with the least instances are the others and sports.

Table 6. Mexican author profiling corpus: distribution of the place of residence trait

Class	Profiles	Average images (α)	Average tweets (α)
North	13	625.23($\pm$442.49)	1594.23($\pm$855.17)
Northwest	80	385.92($\pm$345.95)	1162.17($\pm$866.14)
Northeast	123	460.54($\pm$482.02)	1071.60($\pm$800.66)
Center	191	755.58($\pm$732.74)	1597.83($\pm$922.49)
West	46	611.91($\pm$488.10)	1525.80($\pm$990.62)
Southeast	47	659.12($\pm$732.35)	1284.51($\pm$916.36)

Table 7. Mexican author profiling corpus: distribution of the occupation trait

Class	Profiles	Average images (α)	Average tweets(α)
Arts	38	826.21($\pm$754.71)	1828.23($\pm$834.09)
Student	253	336.57($\pm$259.81)	1184.66($\pm$838.81)
Social	64	1158.15($\pm$867.03)	1362.62($\pm$921.89)
Sciences	25	474.28($\pm$461.97)	1549.64($\pm$947.44)
Sports	12	682.41($\pm$652.27)	1113.00($\pm$892.95)
Administrative	82	894.59($\pm$651.72)	1597.52($\pm$965.65)
Health	15	248.20($\pm$275.05)	1410.20($\pm$1127.04)
Others	11	1026.90($\pm$747.28)	1873.27($\pm$965.63)

3 Analysis of Distributional Term Representations

This section describes a general framework for Author Profiling using distributional term representations (DTRs). Our goal is to overcome, to some extent, the issues naturally inherited by the BoW representation and build instead of a more semantically related representation. Intuitively, DTRs can capture the semantics of a term t_i by exploiting the distributional hypothesis: "words with similar meanings appear in similar contexts". Thus, different DTRs can capture the semantics through the context in different ways and at different levels.

Traditionally, the Author Profiling task has been approached as a single-labeled classification problem, where the different categories (e.g., *male* vs. *female*, or *teenager* vs. *young* vs.

old) stand for the target classes. The common pipeline is as follows: *i*) extracting textual features from the documents; *ii*) building the documents' representation using the extracted features, and *iii*) learning a classification model from the built representations[5].

As it is possible to imagine, extracting the relevant features is a key aspect for learning the textual patterns of the different profiles. Accordingly, previous research has evaluated the importance of thematic (content-based) features [20, 37] and stylistic characteristics [8].

More recently, some works have also considered learning such representations utilizing Convolutional and Recurrent Neural Networks [43, 18, 45].

Although many textual features have been used and proposed, a common conclusion among previous research is that content-based features are the most relevant for this task. The latter can be confirmed by reviewing the results from the PAN[3] competitions [39], where the best-performing systems employed content-based features for representing documents regardless of their genre. This result is somehow intuitive since AP is not focused on distinguishing a particular author through modeling his/her writing style, but on characterizing a group of authors.

The idea is to enrich representations that help to overcome the small-length and high-sparsity issues of social media documents by considering contextual information computed from document occurrence and term co-occurrence statistics. Mainly, we proposed a family of distributional representations based on second-order attributes that allow capturing the relationships between terms and profiles and sub-profiles [29].

These representations obtained the best results in the AP tasks at PAN 2013 and PAN 2014 [28]. Also, we evaluated topic-based representations such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) in the AP task [3], obtaining the best performance at the PAN 2015 as well as showing its superiority against a representation based on manually defined topics utilizing LIWC [4].

[3] A set of shared tasks on digital text forensics: http://pan.webis.de/

In this section, we present a thorough analysis of the pertinence of *distributional term representations* (DTRs) for solving the problem of AP in social media. We aim to highlight the advantages and disadvantages of this type of representation in comparison with traditional topic-based representations such as LSA and LDA.

In summary, the main contributions of this section are:

— We introduce a framework for supervised author profiling in social media domains using DTRs. This framework encompasses the extraction of distributional representation terms as well as the construction of the authors' representation by aggregating the representations of the terms from their documents.

— We evaluate for the first time the document-occurrence representation (DOR) and the term co-occurrence representation (TCOR) in the AP task. These are two simple and well-known term representations from distributional semantics [24].

— We present a comparative analysis of several distributional representations, namely DOR, TCOR, SSR, and word2vec, using the proposed framework for AP. Additionally, we compare their performance against the results from classic bag-of-words and topic-based representations.

3.1 Distributional Term Representations

Let us consider words in the vocabulary as the base terms for building the DTR. More formally, let $\mathcal{D} = \{(d_1, y_1), \ldots, (d_n, y_n)\}$ be a training set of $n-$pairs of documents (d_j) and labels/categories $y_i \in \mathcal{C} = \{C_1, \ldots, C_q\}$. Also let $\mathcal{V} = \{t_1, \ldots, t_m\}$ be the collection vocabulary. In this context, DTRs associates each term $t_i \in \mathcal{V}$ with a term vector $\vec{w_i} \in R^r$, i.e., $\vec{w_i} = \langle w_{i,1}, \ldots, w_{i,r} \rangle$. In this notation $w_{i,j}$ indicates the contribution of distributional feature j to the representation of term t_i. This contribution is particular of each DTR and can be computed in a number of ways.

In the following sections we describe in detail each of the DTRs that we selected for this study. The second step consists in building the document representations by using the term vectors. Formally, the representation of document a d_j, the vector $\vec{d_j}$, is obtained by using the expression 1, where the scalar α_i weighs the relevance of term t_i in the document d_j. Although there are several ways to define this weighting, the most widely used approach is the average of the distribution (i.e., α_i is proportional to the number of terms in the document):

$$\vec{d_j} = \sum_{t_i \in d_j} \alpha_i \cdot \mathbf{w_i}. \quad (1)$$

Different ways to define vectors w_i are briefly explained below. For more details of the formal implementation, consult [23, 47, 1, 28]

3.1.1 Document Occurrence Representation

The document occurrence representation (DOR) can be considered the dual of the TF-IDF representation widely used in the Information Retrieval field [23]. DOR is based on the hypothesis that the semantics of a term can be revealed by its distribution of occurrence-statistics over the documents in the corpus. A term t_i that belongs to the vocabulary $\mathcal{V}$ is represented by a vector of weights associated to documents $\vec{w_i} = \langle w_{i,1}, \cdots, w_{i,N} \rangle$ where N is the number of documents in the collection and $0 \leq w_{i,j} \leq 1$ represents the contribution of document d_j.

3.1.2 Term Co-Occurrence Representation

Term Co-Occurrence Representation (TCOR) is based on co-occurrence statistics [23]. The underlying idea is that the semantics of a term t_i can be revealed by the terms that co-occur with it across the documents collection. Here, each term $t_i \in \mathcal{V}$ is represented by a vector of weights $\vec{w_i} = \langle w_{i,1}, \cdots, w_{i,|\mathcal{V}|} \rangle$ where $0 \leq w_{i,j} \leq 1$ represents the contribution of term t_j to the semantic description of t_i.

3.1.3 Word Embeddings: Word2vec

Recently, a prevalent group of related models for producing word embeddings is word2vec [35]. These models are shallow, two-layer neural networks trained to reconstruct the linguistic contexts of words.

Word2vec takes as its input a large corpus of texts and produces a vector space, typically of a few hundreds of dimensions, where each term in the corpus is assigned to a corresponding vector $\vec{w_i}$ in the space. Thus, once the word vectors have been computed and positioned in the vector space, words that share common contexts in the corpus are located close to each other in the space [34].

In our experiments, we built the word embeddings (i.e., vectors $\vec{w_i}$) using the skip-gram model.

3.1.4 Subprofile Specific Representation

The intuitive idea of the second order attributes consists in representing the terms by their relation with each target class [26, 29]. This can be done by exploiting occurrence-statistics over the set of documents in each one of the target classes.

In this way, we represent each term $t_i \in \mathcal{V}$ with a vector $\vec{w_i} = \langle w_{i,1}, \cdots, w_{i,q} \rangle$, where the scalar $w_{i,k}$ is the degree of association between word t_i and class C_k. Under this DTR, the weight $w_{i,k}$ is directly related to the number of occurrences of term t_i in documents that are labeled with class C_k.

In [29], second order attributes were modeled at sub-profile level; mainly, it was proposed to cluster the instances from each target in order to generate several subclasses. The idea was to consider the high heterogeneity of social media users.

Utilizing this process, the set of target classes C will now correspond to the set of all subgroups from the original target classes. This new representation is called *Subprofile-based Representation* (SSR), and is considered one of the state-of-the-art representations for AP.

3.2 Experiments and Results

This section explains the experiments that were carried out using the proposed framework. As we have previously mentioned, we aim at determining the pertinence of distributional term representations (DTRs) to the AP task in distinct social media domains. Accordingly, this section is organized as follows: first, Subsection 3.2.1 explains the experimental settings for all the experiments, then, Subsection 3.3 describes the results obtained by each DTR in the four different social media domains.

3.2.1 Experimental Setup

Preprocessing: For computing the DTRs of each social media domain we considered the 10,000 most frequent terms. We did not remove any term, i.e., we preserved all content words, stop words, emoticons, punctuations marks, etc. In one previous work [29] demonstrated that preserving only the 10,000 most frequent words is enough for achieving a good representation of the documents.

Text representation: The different DTRs were computed as described in Section 3.1.

Classification: Following the same configuration as in previous works (please refer to [4]), in all the experiments we used the linear Support Vector Machine (SVM) from the LIBLINEAR library with default parameters [11].

Baseline: As baseline we employed the traditional bag-of-words (BoW) representation. We also compared the results from the different DTRs to those obtained by topic modeling representations such as LSA and LDA as well as to those from the top systems from the PAN@2014 AP track.

Evaluation: We performed a stratified 10 cross-fold validation (10-CFV) strategy. For comparison purposes, and following the PAN guidelines, we employed the accuracy as the main evaluation measure. Finally, we evaluated the statistical significance of the obtained results using a 0.05 significance level utilizing the Wilcoxon Signed-Ranks test since is recommended for these cases by [9].

Table 8. F-measure results obtained by the DTRs for the *age* classification problem

Approach	Text genres			
	Blogs	Reviews	Social Media	Twitter
DOR	**0.38**	**0.30**	**0.29**	**0.35**
TCOR	0.22	0.21	0.23	0.31
w2v-wiki	0.21	0.21	0.23	0.30
w2v-sm	0.20	0.20	0.24	0.28
SSR	0.36	0.27	0.26	0.33
Baseline	0.21	0.19	0.23	0.21

Table 9. F-measure results obtained by the employed DTRs for the *gender* classification task

App.	Text genres			
	Blogs	Reviews	Social Media	Twitter
DOR	**0.78***	**0.69***	0.52	0.70
TCOR	0.56	0.62	0.41	0.54
w2v-wiki	0.75*	0.64	0.52	0.69
w2v-sm	0.74	0.64	0.54	0.66
SSR	**0.78***	**0.69***	**0.55***	**0.71**
Baseline	0.72	0.62	0.52	0.70

3.3 Results

This section is organized as follows: first, we show the results from different DTRs for the age and gender classification tasks; then, we compare them against some topic-based representations and the best approaches from PAN 2014.

3.3.1 Age and Gender Identification Using DTRs

Table 8 shows the F-measures results for *age*. Also, Table 9 shows the obtained results for the *gender* classification problems respectively. Each row represents one of the described DTRs, i.e., DOR, TCOR, word2vec, and SSR, while the last row represents the baseline results. Every column refers to a distinct social media genre. In these tables, the best results are highlighted using boldface, and the star symbol ($\star$) indicates the differences that are statistically significant concerning the baseline results (in accordance to the used test; for details refer to Section 3.2.1).

Obtained results indicate that all DTRs, except for TCOR, outperformed the baseline method.

Table 10. Comparison of the best DTRs against topic-based methods in the *age* classification task

Approach	Text genres			
	Blogs	Reviews	Social Media	Twitter
DOR	**0.49**†	0.36†	**0.38**†‡	0.47‡
SSR	0.48†	0.34†	0.37‡	**0.48**†‡
LDA	0.44	0.27	0.37	0.47
LSA	**0.49**	**0.37**	0.36	0.45
[33]	0.38	0.33	0.36	0.44
[48]	0.39	0.31	0.35	0.41
[49]	0.45	0.37	0.42	0.52

Table 11. Comparison of best DTRs against topic-based methods in the *gender* classification task

Approach	Text genres			
	Blogs	Reviews	Social Media	Twitter
DOR	**0.78**†	**0.69**†	0.52	0.70†
SSR	**0.78**†	**0.69**†	**0.55**$^{\dagger\ \ddagger}$	
LDA	0.61	0.55	0.52	0.64
LSA	**0.78**	**0.69**	0.53	0.70
[33]	0.57	0.66	0.53	0.66
[48]	0.64	0.68	0.54	0.51
[49]	0.82	0.71	0.57	0.78

In particular, DOR and SSR show statistically significant differences. These two methods obtained comparable results, being DOR slightly better than SSR in 5 out of 8 classification problems, which is an interesting result since SSR was among the winning approaches at PAN 2014. On the other hand, we attribute the low accuracy results showed by TCOR to the strong expansion that it imposes to the document representations.

Considering direct term co-occurrences causes the inclusion of many unrelated and unimportant terms in the document vectors, and, therefore, it complexities the extraction of profiling patterns.

Finally, another essential aspect to notice is the fact that both *w2v-wiki* and *w2v-sm* obtained similar results in each of the classifications problems, although the former learned the embeddings from a corpus that is not thematically and neither stylistically similar to the social media content. We presume these results could be explained by the relatively small size of the social media training collections, and, at the same time, by the large size and broad coverage of the used Wikipedia dataset, which has a vocabulary of 1,033,013 words.

Tables 10 and 11 compare the results from DOR and SSR, the best DTRs according to the previous results, against the results from two well-known topic-based representations, namely LDA and LSA.

Regarding the LSA results, it is possible to observe, on the one hand, that for *age* classification (refer to Table 10), its average performance is similar to the one from DOR, i.e., 42%. However, the only domain in which LSA outperforms DOR is in the reviews dataset. Nonetheless, there is no significant difference between these results. On the other hand, for *gender* classification (Table 11), LSA was not able to improve any result from DOR and SSR. It is important to mention that, although their results are comparable, LSA is a parametric method, and, therefore, tunning is required.

Finally, the works [33], [48] and [49] are the best results for the PAN@2014 forum.

4 Image Author Profiling Approach

4.1 Open-Vocabulary Method

The adopted UAIA method for labeling images with an open vocabulary approach was proposed in [36]. The general idea of this method relies on the use of a multimodal indexing $\mathcal{M}$ composes of visual prototypes that are used for labeling new images.

Given a reference collection of documents $\mathcal{D}$ that include texts $\mathcal{T}$ and images $\mathcal{I}$. First, each image in $\mathcal{V}$ is represented by a visual feature $\mathbf{v}_i$. In our case, we use the VGG-16 pre-trained model proposed in [46] for visual extraction. Then, each extracted word, i.e. $\mathbf{t}_i$, from $\mathcal{T}$ is represented by visual vector resulting from combining images that co-occur with the word, i.e., visual features of images included in documents where the word appears.

Mathematically, multimodal indexing could be done as follows:

$$\mathcal{M} = \mathcal{T}^T \cdot \mathcal{V}, \tag{2}$$

ISSN 2007-9737

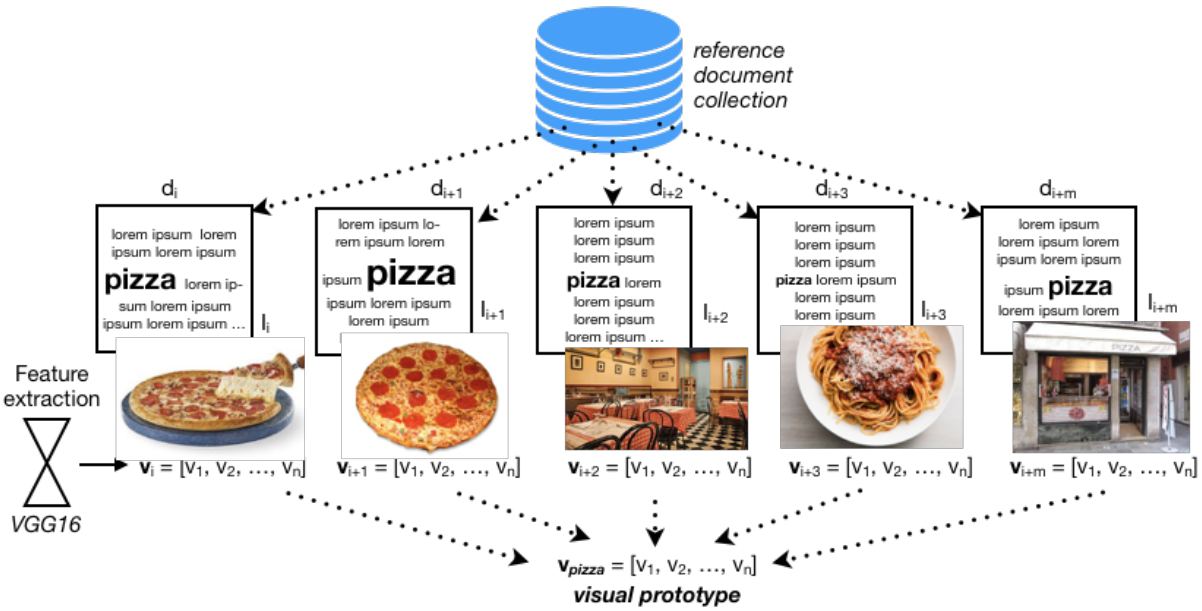

Fig. 2. Illustrate of how the annotator builds visual prototypes. In this example the method uses all images where appears the word 'pizza' and as result obtains its visual prototype

where $\mathcal{M}$ is the multimodal indexing obtained by the product of textual $\mathcal{T}$ and visual features $\mathcal{V}$ of documents. The advantage of the UAIA method is the capability to build visual prototypes from free and large vocabularies extracted from reference collections. The whole process is illustrated in Figure 2, for the case, note that the word 'pizza' is bigger when it appears with more frequency in the document. This mechanism prevents to add images with no relevance to the word.

For annotating a new image, first, it is described in a common representation to the visual prototypes, then it is compared for estimating a similarity score based on cosine distance:

$$cosine(\mathbf{q}, \mathcal{M}) = \frac{\mathbf{q} \cdot \mathbf{m}_i}{|\mathbf{q}| \times |\mathbf{m}_i|}, \quad (3)$$

where $\mathbf{q}$ is the visual representation of the query image, and $\mathbf{m}_i$ is the $i-th$ visual prototype in $\mathcal{M}$. The query image is compared with each visual prototype in $\mathcal{M}$, and n of the most similar visual prototypes, that is, the n of the most similar words are used for annotating the image.

LSA captures the topics in a corpus applying a mathematical technique called singular value decomposition (SVD) while preserving the similarity structure among the texts. The underlying idea is that the aggregate of all the word contexts, in which a given word does and does not appear, it provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. Unlike BoW, LSA represents each document D_j into a k-dimensional vector where k represents the number of discovered topics, thus, $D_j = \{x_1, x_2, ..., x_i, ..., x_k\}$.

Each dimension i in the vector represents the weight of a topic i in the document j [16]. For the performed experiments, when we apply LSA on the posts' text we express it as LSA$_T$, and when we use LSA on the labels extracted from the image annotation methods, we express it as LSA$_I$.

4.2 Results

Table 12 shows the results of the proposal compared with the individual text results as BoW and LSA$_T$ for the Mexican collections for the 3 traits. Also, we compared the proposal results with the AlexNet [22] and RCNN [14] models. These models are based on deep learning and represent

Table 12. Obtained performance for the gender, occupation, and location tasks on the MEX-A3T corpus

Gender Identification Task			
Approach		*Accuracy*	*F1*
Textual baselines	BoW	0.80	0.80
	LSA_T	0.79	0.79
Visual baselines	AlexNet [22]	0.65	0.65
	RCNN [14]	0.64	0.64
Proposed	BoL	0.74*	0.74
	LSA_I	**0.79***	**0.79**
Multi modal	DOR+LSA_I	**0.79***	**0.79**
Occupation Identification Task			
Textual baselines	BoWk	0.64	0.34
	LSA_T	0.65	0.25
Visual baselines	AlexNet [22]	0.52	0.23
	RCNN [14]	0.54	0.24
Proposed	BoL	0.63*	**0.34**
	LSA_I	**0.65***	**0.34**
Multi modal	DOR+LSA_I	**0.68***	**0.39**
Location Identification Task			
Textual baselines	BoW	0.52	0.37
	LSA_T	0.71	0.57
Visual baselines	AlexNet [22]	0.35	0.24
	RCNN [14]	0.35	0.23
Proposed	BoL	0.44*	0.28
	LSA_I(k=100)	**0.50***	**0.31**
Multi modal	DOR+LSA_I	**0.68***	**0.58**

each image as a vector of 1000 semantic features. The proposal is represented as BoL for the bag of labels and LSA_I for the implementation of LSA for the labels of BoL.

For all traits, the textual representations obtain better or very similar results than the image representations. This indicates that the textual information is more valuable than the image information. Also, for the three traits, the proposal results overcome the deep learning based methods. Particularly, LSI_I overcomes the BoL implementation, it seems that to group the labels by their contest provides a better representation.

Finally, we implement a fusion strategy for taking advantage of both modalities. We use the late fusion [30] concatenated both spaces, DOR (the best DTR result) and LSA_I (the best image representation). As we can see, the most noticeable difference occurs for the location trait

however, for occupation, the results are better with the fusion too. But, for gender trait, the text result is still the best result. This could occurs because the occupation and location are traits unbalanced and the gender trait is balanced, Thus the occupation and location traits are harder tasks than the gender classification and it is necessary more information to help the classification [31].

5 Cross-Language Gender Prediction

We appraise the robustness of our proposed method under a cross-lingual scenario [12, 27]. For this, we performed several experiments training and evaluating using distinct *source* and *target* languages, and compared against the best results obtained in a monolingual situation.

The hypothesis behind this idea establishes that users with distinct native languages, having a similar profile, will share analogous images. To the best of our knowledge, this is the first attempt in proposing a cross-language gender prediction method based on merely visual information.

In order to prove this hypothesis, we performed a series of experiments for the gender prediction task[4]. Similar to the previous experiments, we compare the performance of the closed vocabulary approaches (AlexNet and RCNN) against the performance of the open vocabulary approach (LSA$_I$) under a cross-lingual scenario.

Table 13. Cross language results using AlexNet as image annotation method

Source language	*Target language*	*Acc.*	F_1	F_1 Male	F_1 Female
EN	EN	0.58	0.58	0.56	0.60
SP	EN	0.59	0.59	0.59	0.59
SP+EN	EN	**0.61**	**0.61***	0.60	0.62
SP	SP	0.65	0.65	0.65	0.65
EN	SP	0.64	0.64	0.64	0.64
SP+EN	SP	**0.66**	**0.66***	0.66	0.66

[4]The *gender* trait is the only common trait among both datasets, i.e., PAN 2014 and MEX-A3T

Table 14. Cross language results using RCNN as image annotation method

Source language	*Target language*	*Acc.*	F_1	F_1 Male	F_1 Female
EN	EN	0.56	0.56	0.56	0.56
SP	EN	0.60	0.55	0.70	0.40
SP+EN	EN	**0.61**	**0.61***	0.61	0.61
SP	SP	0.64	0.64	0.64	0.64
EN	SP	0.64	0.59	0.46	0.72
SP+EN	SP	**0.65**	**0.65***	0.64	0.66

Table 15. Cross language results using the proposed method under the LSA$_I$. The number between parenthesis indicates the value of the k parameter for the LSA method

Source language	*Target language*	*Acc.*	F_1	F_1 Male	F_1 Female
EN	EN(100)	0.72	0.72	0.71	0.72
SP	EN(100)	0.60	0.55	0.70	0.40
SP+EN	EN(50)	**0.84**	**0.84***	0.84	0.84
SP	SP(100)	0.79	0.79	0.79	0.79
EN	SP(100)	0.64	0.59	0.46	0.72
SP+EN	SP(50)	**0.80**	**0.80***	0.80	0.80

5.1 Results

Table 13, 14 and 15 show the obtained results using ALexNet, RCNN and LSA$_I$ methods for labeling the visual information. It is interesting to observe that when only one language is used for training (EN$\rightarrow$ EN, SP$\rightarrow$EN, SP$\rightarrow$SP, EN$\rightarrow$SP), achieved performance is very similar for all AIA methods. However, a significant improvement is obtained when the combination of the two languages (SP+EN) is employed to train the classification model Particularly, observe the LSA$_I$ method (Table 15) outperforms both AlexNet (Table 13) and RCNN (Table 14) configurations.

These results evidentiate that similar users share in fact similar images, allowing an automatic classifier to distinguish among users, regardless of their native language. In order to exemplify this affirmation, we took on the task of retrieving the most important images from the top 5 topics

(a) Topic 1: People **(b)** Topic 2: Sports

(c) Topic 3: Women **(d)** Topi 4: Technology

(e) Topic 5: Diagrams

Fig. 3. Representative images for each topic extracted with LSA

identified by the LSA_I approach. Figure 3 illustrates the retrieved images.

For each topic six images are shown, where the three from the left correspond to Spanish speaking users, and the three on the right to English speaking users. After observing the retrieved images, is possible to conclude that shared images by users not sharing language, at least between males and females, contain similar characteristics. This language and culture independent phenomenon indicates that is possible to configure cross-lingual AP methods.

6 Conclusions

As a result of this work, the following conclusions were obtained.

DTR's have advantages in the author profiling task compared with other approaches to capture the content of the texts. In particular, DOR presents the best behavior, besides that DOR is not a parameterized approach, which causes it to be a simpler and more efficient approach to this task. Also, a significant advantage of DOR is its robustness across different social media genres, contrary to others approaches.

Automatic image annotation based on open vocabulary approaches is better to represent the images than the closed vocabulary approaches for the Author profiling task.

There is complementarity among the textual and image modalities since it is possible to overcome the individual results with fusion schemes.

Also, it is possible to use image information from another corpus, even if the corpus is in another language. This seems reasonable, taking into account that images are language independent. It seems that the open vocabulary approach with LSA_I represents better the images from different native languages users.

Acknowledgment

Álvarez-Carmona thanks for doctoral scholarship CONACyT-Mexico 401887.

References

1. **Álvarez-Carmona, M. Á. (2019).** *Author profiling in social media with multimodal information*. Ph.D. thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica.
2. **Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Reyes-Meza, V., & Rico-Sulayes, A. (2018).** Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. *Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September*.
3. **Álvarez-Carmona, M. A., López-Monroy, A. P., Montes-y Gómez, M., Villaseñor-Pineda, L., & Escalante, H. J. (2015).** INAOE's participation at PAN'15: Author profiling task. *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, Vol. 1391.
4. **Álvarez-Carmona, M. A., López-Monroy, A. P., Montes-y Gómez, M., Villaseñor-Pineda, L., & Meza, I. (2016).** Evaluating topic-based representations for author profiling in social media. *Ibero-American Conference on Artificial Intelligence*, Springer, pp. 151–162.
5. **Alvarez-Carmona, M. A., Villatoro-Tello, E., Villasenor-Pineda, L., et al. (2019).** A comparative analysis of distributional term representations for author profiling in social media. *Journal of Intelligent & Fuzzy Systems*, Vol. 36, No. 5, pp. 4857–4868.
6. **Aragón, M. E., Álvarez-Carmona, M. Á., Montes-y Gómez, M., Escalante, H. J., Villasenor-Pineda, L., & Moctezuma, D. (2019).** Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. *Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain*.
7. **Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. W. (2005).** Lexical predictors of personality type. *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*.
8. **Bergsma, S., Post, M., & Yarowsky, D. (2012).** Stylometric analysis of scientific articles. *Proceedings of the 2012 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, pp. 327–337.

9. **Demšar, J. (2006).** Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, Vol. 7, No. Jan, pp. 1–30.
10. **Eftekhar, A., Fullwood, C., & Morris, N. (2014).** Capturing personality from Facebook photos and photo-related activities: How much exposure do you need? *Computers in Human Behavior*, Vol. 37, pp. 162–170.
11. **Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008).** Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874.
12. **Feliciano-Avelino, I., Álvarez-Carmona, M. Á., Escalante, H. J., Montes-y Gómez, M., & Villaseñor-Pineda, L. (2019).** Cross-cultural image-based author profiling in twitter. *Mexican International Conference on Artificial Intelligence*, Springer, pp. 353–363.
13. **Gelbukh, A. (2019).** Computational linguistics: Introduction to the thematic issue. *Computación y Sistemas*, Vol. 23, No. 3.
14. **Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014).** Rich feature hierarchies for accurate object detection and semantic segmentation. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587.
15. **Grimshaw, M. (2013).** *The Oxford handbook of virtuality*. Oxford University Press.
16. **Hofmann, T. (1999).** Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 50–57.
17. **Kharroub, T. & Bas, O. (2015).** Social media and protests: An examination of Twitter images of the 2011 Egyptian revolution. *New Media & Society*, pp. 1461444815571914.
18. **Kodiyan, D., Hardegger, F., Neuhaus, S., & Cieliebak, M. (2017).** Author profiling with bidirectional rnns using attention with grus, pp. 1–10.
19. **Koppel, M., Akiva, N., Alshech, E., & Bar, K. (2009).** Automatically classifying documents by ideological and organizational affiliation. *Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on*, IEEE, pp. 176–178.
20. **Koppel, M., Argamon, S., & Shimoni, A. R. (2002).** Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, Vol. 17, No. 4, pp. 401–412.
21. **Koppel, M., Schler, J., & Zigdon, K. (2005).** Determining an author's native language by mining a text for errors. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, ACM, pp. 624–628.
22. **Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012).** Imagenet classification with deep convolutional neural networks. In **Pereira, F., Burges, C. J. C., Bottou, L., & Weinberger, K. Q.**, editors, *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., pp. 1097–1105.
23. **Lavelli, A., Sebastiani, F., & Zanoli, R. (2004).** Distributional term representations: an experimental comparison. *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, ACM, pp. 615–624.
24. **Lavelli, A., Sebastiani, F., & Zanoli, R. (2004).** Distributional term representations: An experimental comparison. *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, ACM, New York, NY, USA, pp. 615–624.
25. **Levy, O., Goldberg, Y., & Dagan, I. (2015).** Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, Vol. 3, pp. 211–225.
26. **Li, Z., Xiong, Z., Zhang, Y., Liu, C., & Li, K. (2011).** Fast text categorization using concise semantic analysis. *Pattern Recognition Letters*, Vol. 32, No. 3, pp. 441–448.
27. **López, R., Peñaloza, D., Beingolea, F., Tenorio, J., & Sobrevilla Cabezudo, M. (2019).** An exploratory study of the use of senses, syntax and cross-linguistic information for subjectivity detection in spanish. *Computación y Sistemas*, Vol. 23, No. 3.
28. **López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., & Villaseñor-Pineda, L. (2014).** Using intra-profile information for author profiling. *CLEF 2014 Working Notes*, pp. 1116–1120.
29. **López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., & Stamatatos, E. (2015).** Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems*, Vol. 89, pp. 134–147.

30. **Loyola-González, O., López-Cuevas, A., Medina-Pérez, M. A., Camiña, B., Ramírez-Márquez, J. E., & Monroy, R. (2019).** Fusing pattern discovery and visual analytics approaches in tweet propagation. *Information Fusion*, Vol. 46, pp. 91–101.
31. **Loyola-González, O., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., & García-Borroto, M. (2016).** Effect of class imbalance on quality measures for contrast patterns: An experimental study. *Information Sciences*, Vol. 374, pp. 179–192.
32. **Loyola-González, O., Medina-Pérez, M. A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., Monroy, R., & García-Borroto, M. (2017).** Pbc4cip: A new contrast pattern-based classifier for class imbalance problems. *Knowledge-Based Systems*, Vol. 115, pp. 100–109.
33. **Maharjan, S., Shrestha, P., & Solorio, T. (2014).** A simple approach to author profiling in mapreduce. *CLEF (Working Notes)*, pp. 1121–1128.
34. **Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013).** Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
35. **Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013).** Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pp. 3111–3119.
36. **Pellegrin, L., Escalante, H. J., Montes-y Gómez, M., & González, F. A. (2016).** Local and global approaches for unsupervised image annotation. *Multimedia Tools and Applications*, Vol. 76, No. 15, pp. 16389–16414.
37. **Poulston, A., Waseem, Z., & Stevenson, M. (2017).** Using tf-idf n-gram and word embedding cluster ensembles for author profiling, pp. 1–6.
38. **Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., & Daelemans, W. (2014).** Overview of the 2nd author profiling task at pan 2014. *Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes)*, pp. 1–30.
39. **Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., & Stein, B. (2018).** Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter. *Working Notes Papers of the CLEF*, pp. 1–38.
40. **Rangel, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015).** Overview of the 3rd author profiling task at pan 2015. *CLEF*, sn, pp. 2015.
41. **Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006).** Effects of age and gender on blogging. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pp. 199–205.
42. **Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., Park, G. J., Lakshmikanth, S. K., Jha, S., Seligman, M. E., et al. (2013).** Characterizing geographic variation in well-being using tweets. *ICWSM*, pp. 583–591.
43. **Sierra, S. & González, F. A. (2018).** Combining textual and visual representations for multimodal author profiling. *Working Notes Papers of the CLEF*, Vol. 2125, pp. 219–228.
44. **Skalmowski, W. (2016).** Review of harris, zellig (1968) mathematical structures of language. *ITL-International Journal of Applied Linguistics*, Vol. 4, No. 1, pp. 56–61.
45. **Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., & Ohkuma, T. (2018).** Text and image synergy with feature cross technique for gender identification. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, volume 2125, pp. 10–22.
46. **Tindall, L., Luong, C., & Saad, A. (2015).** Plankton classification using vgg16 network.
47. **Villegas, M. P., Garciarena Ucelay, M. J., Fernández, J. P., Álvarez Carmona, M. A., Errecalde, M. L., & Cagnina, L. (2016).** Vector-based word representations for sentiment analysis: a comparative study. *XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016).*
48. **Villena Román, J. & González Cristóbal, J. C. (2014).** Daedalus at pan 2014: Guessing tweet author's gender and age, pp. 1157–1163.
49. **Weren, E. R., Kauer, A. U., Mizusaki, L., Moreira, V. P., de Oliveira, J. P. M., & Wives, L. K. (2014).** Examining multiple features for author profiling. *Journal of Information and Data Management*, Vol. 5, No. 3, pp. 266.
50. **Wu, Y.-C. J., Chang, W.-H., & Yuan, C.-H. (2014).** Do Facebook profile pictures reflect user's personality? *Computers in Human Behavior*, Vol. 51, pp. 880–889.

Article received on 14/06/2020; accepted on 21/07/2020.
Corresponding author is Manuel Montes y Gómez.