Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter

Francisco Rangel[1,2] Paolo Rosso[2] Manuel Montes-y-Gómez[3]
Martin Potthast[4] Benno Stein[5]

[1]Autoritas Consulting, S.A., Spain
[2]PRHLT Research Center, Universitat Politècnica de València, Spain
[3]INAOE, Mexico
[4]Leipzig University, Germany
[5]Web Technology & Information Systems, Bauhaus-Universität Weimar, Germany

pan@webis.de http://pan.webis.de

Abstract This overview presents the framework and the results of the Author Profiling shared task at PAN 2018. The objective of this year's task is to address gender identification from a multimodal perspective, where not only texts but also images are given. For this purpose a corpus with Twitter data has been provided, covering the languages Arabic, English, and Spanish. Altogether, the approaches of 23 participants are evaluated.

1 Introduction

Author profiling is the analysis of shared content in order to predict different attributes of authors such as gender, age, personality, native language, or political orientation. Supported by the huge amount of information that is available on social media platforms, author profiling has gained a lot of interest. Being able to infer an author's gender, age, native language, dialects, or personality opens a world of possibilities—among others in marketing, where companies may analyze online reviews to improve targeted advertising, or in forensics, where the profile of authors could be used as valuable additional evidence in criminal investigations, and in security, where knowing the demographics of social media users (age and gender), as well as cultural and social context such as native language and dialects, may help to identify potential terrorists [50].

In the following we provide a historical outline of previous editions of this task. In the Author Profiling task at PAN 2013[1] [44], the identification of age and gender relied on a large corpus collected from social media, both for English and Spanish. In PAN 2014[2] [45], we continued focusing on age and gender aspects but, in addition, compiled a corpus of four different genres, namely social media, blogs, Twitter, and hotel reviews. Except for the hotel review subcorpus, which was available for English only, all documents were provided in both English and Spanish. Note that most of the

[1] http://webis.de/research/events/pan-13/pan13-web/author-profiling.html
[2] http://webis.de/research/events/pan-14/pan14-web/author-profiling.html

existing research in computational linguistics [6] and social psychology [39] focuses on the English language, and the question is whether the observed relations pertain to other languages and genres as well. In this vein, in PAN 2015[3] [46], we included two new languages, Italian and Dutch, besides a new subtask on personality recognition in Twitter. In PAN 2016[4] [49], we investigated the effect of cross-genre information: the models are trained on a certain genre (here: Twitter) and evaluated on another genre different than Twitter. In PAN 2017[5] [18], we considered the language variety identification together with the gender dimension. We evaluated this new subtask in four languages: Arabic, English, Portuguese and Spanish.

Social media data cover a wide range of modalities such as text, images, audio, and video, all of which containing useful information to be exploited for extracting valuable insights from users. Consequently, the objective of this year's evaluation[6] is to address gender identification from a multimodal perspective: not only texts but also images are given. For this purpose a corpus with Twitter data has been provided, covering the languages: Arabic, English, and Spanish.

The remainder of this paper is organized as follows. Section 2 covers the state of the art, Section 3 describes the corpus and the evaluation measures, and Section 4 presents the approaches submitted by the participants. Sections 5 and 6 discuss results and draw conclusions respectively.

2 Related Work

The relationship between personal traits and the use of language has been widely studied by the psycholinguistics Pennebaker [40]. He analysed how the use of the language varies depending on personal traits. For example, in regards to the authors' gender, he found out that in English women use more negations or first persons, because they are more self-concientious, whereas men use more prepositions in order to describe their environment. These finding are the basis of LIWC (Linguistic Inquiery and Word Count) [39] that is one of the most used tools in author profiling.

Initial investigations in author profiling [6, 25, 13, 27, 53] focused mainly on formal texts and blogs. Their reported accuracies ranged from 75% to 80%. Nevertheless, nowadays researchers focused mainly on social media, where the language is more spontaneous and less formal. It should be highlighted the contribution of different researchers that used the PAN datasets. For example, the authors in [34] showed how to deal with a large dataset such as the PAN-AP-2013 with 3 million features with a MapReduce configuration. With the same dataset, the authors in [66] showed the contribution of information retrieval-based features. Following Pennebaker findings about the relationship between emotions and gender, the authors in [43] proposed the EmoGraph graph-based approach to capture how users convey verbal emotions in the morphosyntactic structure of the discourse and showed competitive results with the best performing systems at PAN-2013 and demonstrating the robustness of the approach against

[3] http://pan.webis.de/clef15/pan15-web/author-profiling.html
[4] http://pan.webis.de/clef16/pan16-web/author-profiling.html
[5] http://pan.webis.de/clef17/pan17-web/author-profiling.html
[6] https://pan.webis.de/clef18/pan18-web/author-profiling.html

genres and languages at PAN-2014 [42]. Recently, Bayot and Gonçalves [10] used the PAN-AP-2016 dataset to show that word embeddings worked better in case of gender identification than TF-IDF. Finally, it is worth mentioning the second order representation based on relationships between documents and profiles used by the best performing team in three editions of PAN [29, 30, 4], as well as the performance of the combination of n-grams as shown by the authors [9] of the best performing team at PAN 2017.

The investigation in Arabic is more scarce and most of the research focused on other genres than social media. For example, Estival *et al.* [17] focused on Arabic emails. The authors reported accuracies of 72.10%. Similarly, Alsmearat *et al.* [2] focused on Arabic newsletters. They initially reported an accuracy of 86.4% that was increased to 94% in an extension of their work [1]. With respect to social media, AlSukhni & Alequr [3] focused on Arabic tweets and they reported accuracies of 99.50%. They improved a bag-of-words model with the use of the Twitter authors' names.

The use of visual features for author profiling has been less studied. A common approach for gender identification is the use of frontal facial images [36, 59, 16]. The authors in [36] trained SVM with 1,755 low resolution thumbnail faces (21x12 pixels) from the FERET face database[7] obtaining an error of 3.4%. The authors in [59] used Principal Component Analysis to represent each image in a smaller dimensional space, reducing the error from 17.7% to 11.3% with a neural network. The authors in [16] experimented with 120 combinations of automatic face detection, face alignment and gender classification. They found out that the automatic face alignment did not increase the gender classification rates, whereas the manual alignment did. The authors evaluated several machine learning algorithms, obtaining the best results with SVM. They also saw that the classification did not depend on the size of the images. Recently, user annotated data have been used more and more. For example, Twitter has been used as repository to learn and evaluate gender identification systems. In this sense, the authors in [33] used automatic image annotations and the authors in [55] proposed a Multi-task Bilinear Model to combine the visual concept detector with the feature extractor to predict gender in Twitter. Similarly, the authors in [8] used 56 image aesthetic features to gender identification in 24,000 images provided by 120 FlickR users, obtaining 82.50% of accuracy.

3 Evaluation Framework

The purpose of this section is to introduce the technical background. We outline the construction of the corpus, introduce the performance measures and baselines, and describe the idea of so-called software submissions.

3.1 Corpus

The focus of this year's task is on gender identification in Twitter from a multimodal perspective: besides textual information, the participants are provided also with images. The task is framed as a multilingual task, covering the languages Arabic, English, and Spanish.

[7] https://www.nist.gov/programs-projects/face-recognition-technology-feret

Table 1. Number of authors per language and subset. The corpus is balanced regarding gender and contains 100 tweets and 10 images per author.

	(AR) Arabic	(EN) English	(ES) Spanish	Total
Training	1,500	3,000	3,000	7,500
Test	1,000	1,900	2,200	5,100
Total	2,500	4,900	5,200	12,600

The PAN-AP-2018 corpus is based on the PAN-AP-2017 corpus [48], extended by images that have been shared in the respective Twitter timelines. More specifically, PAN-AP-2018 contains those authors from the PAN-AP-2017 corpus who still have a Twitter account and who have shared at least 10 images. Table 1 overviews the key figures of the corpus. Moreover, the corpus is balanced with regard to gender and it contains 100 tweets per author.

3.2 Performance Measures

The participants were asked to submit per author three predictions according to the following *modalities*: *a)* text-based, *b)* image-based, and *c)* a combination of both. It was allowed to approach the task in a favoured language and a favoured modality; however, we encouraged them to participate in all languages and all modalities.[8]

For each language and for each modality the accuracy was computed. Note that the accuracy of the combined approach has been chosen as overall accuracy for the given language; if only the textual approach was submitted, its accuracy has been used. The final ranking has been calculated as the average accuracy per language as defined by the following equation:

$$ranking = \frac{acc_{ar} + acc_{en} + acc_{es}}{3} \tag{1}$$

3.3 Baselines

In order to assess the complexity of the subtasks per language and to compare the performances of the participants approaches, we propose the following baselines:

- *BASELINE-stat.* A statistical baseline that emulates random choice. As there are two classes and the number of instances is balanced, the random choice baseline is 50% accuracy. This baseline applies for both modalities, images and texts.
- *BASELINE-bow.* To approach the textual modality, we have represented the documents under a bag-of-words model with the 5,000 most common words in the training set, weighted by absolute frequency. The texts are preprocessed as follows: lowercase words, removal of punctuation signs and numbers, and removal of stop words for the corresponding language.

[8] From the 23 participants, 22 participated in the Arabic and Spanish tasks, and all of them in the English tasks. All of them approached the task with text features, where 12 participants also used images.

– *BASELINE-rgb.* To approach the image modality, we represent the photos as follows. For each author, we obtain the RGB color for each pixel in his/her photos. We represent the author with the following descriptive statistics of the RGB values: minimum, maximum, mean, median, and standard deviation.

3.4 Software Submissions

We asked for software submissions (as opposed to run submissions). Within software submissions, participants submit executables of their author profiling softwares instead of just the output (also called "run") of their softwares on a given test set. Our rationale to do so is to increase the sustainability of our shared task and to allow for the re-evaluation of approaches to Author Profiling later on, and, in particular, on future evaluation corpora. To facilitate software submissions, we develop the TIRA experimentation platform [20, 21], which renders the handling of software submissions at scale as simple as handling run submissions. Using TIRA, participants deploy their software on virtual machines at our site, which allows us to keep them in a running state [22].

4 Overview of the Submitted Approaches

This year, 23 teams participated in the Author Profiling shared task and 22 of them submitted the notebook paper.[9] We analyse their approaches from three perspectives: preprocessing, features to represent the authors' texts, and classification approaches.

4.1 Preprocessing

Various participants cleaned the textual contents to obtain plain text. Most of them removed or normalised Twitter-specific elements such as URLs, user mentions, or hashtags [14, 60, 58, 41, 52, 23, 65, 35, 64, 37, 28]. Some participants also lowercased the words [65, 64, 37, 11, 28, 58, 52, 23]. The authors in [14, 58, 23, 64] removed punctuation signs; character flooding has been removed by the authors in [14, 41]. Stopwords have been removed by the authors in [14, 41, 23, 64], and contractions and abbreviations have been expanded by the authors in [58, 41]. The authors in [14] applied specific preprocessing to Arabic texts, such as normalisation and diacritics removal.

Only three participants preprocessed images. The authors in [60] applied direct resizing and resizing with cropping, as well as normalisation by subtracting the average RGB value per language. The authors in [35] rescaled all images to 64x64 and used only those containing human faces, while the authors in [56] rescaled all images to 224 pixel width, maintaining the aspect ratio.

4.2 Features

In previous editions of the author profiling task at PAN as well as in the referred literature, features used for representing text documents have been distinguished as either

[9] Hacohen-Kerner *et al.* described in their working note the participation of two teams.

content-based or style-based. However, this year several participants have employed deep learning techniques. It is interesting to differentiate among traditional features and these new methods in order to compare their performance in the author profiling task. While the authors in [35, 64, 11, 32, 60] represented documents with word embeddings, the authors in [52] used character embeddings. Moreover, the authors in [58, 51, 32] also used traditional features such as character, word, and/or POS n-grams. The authors in [38] combined word embeddings for English as well as stylistic features; however, for Spanish and Arabic they used LSA instead of word embeddings.

Traditional features such as character and word n-grams have been widely used [65, 61, 37, 28, 15, 23, 58, 14]. Style features have been also used by some participants [38, 26, 23]. For example, the authors in [38] used the counts of stopwords, punctuation marks, emoticons, and slang words (only for English). The authors in [26] combined POS tags n-grams with syntactic dependencies to model the use of amplifiers, verbal constructions, pronouns, subjects and objects, types of adverbials, as well as the use of interjections and profanity. The authors in [23] counted the average number of characters and the average number of words per tweet. The authors in [65] also used emojis, whereas the authors in [19] used only the skewness calculated from a variation of the Low Dimensionality Statistical Embedding (LDSE) [47]. The authors in [5] combined ensembles of word and character n-grams with bag-of-terms and second order features [29, 30, 31], which relates documents with authors' profiles.

With respect to the representation of images several approaches have been presented. For example, some participants tried to detect faces in images [58, 14, 64]. In this regard, the authors in [64] used face vectors from images that contained only faces. Besides faces the authors in [14] detected also objects and quantified local binary patterns and color histograms. Other authors used image resources, such as [38], who applied an image captioning system [63]. Similarly, the authors in [37] used a known image feature extraction tool [7] to obtain features about the number of faces in the images, as well as the expressed emotions or their gender. The authors in [5] used ImageNet [57] to obtain VGG16[10] features, and the authors in [52] built a language-independent model with TorchVision.[11] The authors in [60] also used a pre-trained Convolutional Neural Network (CNN) based on VGG16. Other participants approached the task with their own set of features, such as the authors in [23] who combined three sets of characteristics: Shift, RGB histogram, and VGG. The authors in [61] designed a variant of the Bag-of-Visual-Words (BoVW) by using the DAISY [62] feature descriptor and encoded the images by the set of visual words.

4.3 Classification Approaches

Regarding the deep learning approaches, the authors with the overall highest accuracy [60] used Recurrent Neural Networks (RNN) for texts and CNN for images. CNNs have also been used by the authors in [5, 52, 54, 35], while RNNs have also been used by the authors in [11]. Interestingly, the authors in [52] used CNN only for texts and ResNet18 [24] for images. In the same vein, the authors in [64] approached the images

[10] Visual Geometry Group: http://www.robots.ox.ac.uk/~vgg/research/very_deep
[11] https://pytorch.org/docs/stable/torchvision/index.html

with SVM but used Bi-LSTM for texts. The authors in [58] used CNN for images and an ensemble of Naive Bayes and RNN for texts. Finally, the authors in [41] approached the task with dense neural networks.

Some participants still used traditional machine learning algorithms such as logistic regression [51, 23, 65, 37], SVMs [32, 5, 14, 38, 61, 64], multilayer perceptron [23], a basic feed-forward network [28], and distance-based methods [61, 26]. It is worth to mention the approach in [19], who used a simple IF condition with respect to only one feature, allowing the system to process the whole dataset in seconds while achieving a decent performance.

5 Evaluation and Discussion of the Submitted Approaches

Although we encouraged to consider both modalities, some participants approached the problem with text features only. We present the results separately to account for this fact.

5.1 Gender Identification with Text Features

As can be seen in Table 2, the best results were obtained for English (82.21%) [15] and Spanish (82%) [15], although being only slightly better than for Arabic (81.70%) [61]. This similarity is also reflected by the mean accuracies, which are 74.85% for Arabic, 76.93% for English, and 75.46% for Spanish. Taking a closer look at the distributions (Figure 1) shows a different characteristic for English: the median is higher and approximately equal to the Q3 of the other languages, while the interquartile range is smaller. The similarity in the mean value is due to the two outliers (55.21% [26] and 66.580% [51]). This fact is highlighted in the density chart (Figure 2), where the curve for the English language is more skewed to the right and the kurtosis is higher since there are more results concentrated around 80%.

The best result for Arabic (81,70%) is from the authors in [61]; they performed several preprocessing steps and trained an SVM with word n-grams, character n-grams, and skip-grams of different lengths and different weighing schemes such as boolean, tf, and tf-idf. There is no statistical significance with respect to the second (81.20%) [56] and third (80.90%) [15] best results. The authors approached the task with character n-grams and combinations of different types of n-grams. The best result for English (82.21%) comes from the authors in [15]. There is no statistical significance with the second (81.21%) [61] and third (81.16%) [37] best results. The authors in [37] used Logistic Regression with word and character n-grams. Finally, for Spanish, the best result (82%) is from the authors in [15]. Again, there is no statistical significance regarding the second (80.36%) [64] and third (80.27%) [37] best systems. The authors in [64] used a bi-LSTM with pre-trained word embeddings.

With respect to the provided baselines, we can discard the statistical one since its results are much lower than those obtained by the participants. The BOW baseline is at rank 17 out of 22 in the overall ranking.[12] Furthermore, for Arabic the obtained result

[12] The system of Kalgren *et al.* is not count since they participated in the English tasks only.

Table 2. Accuracy per language in the gender identification task with text features.

Ranking	Team	Arabic	English	Spanish	Average
1	Daneshvar	0.8090	**0.8221**	**0.8200**	**0.8170**
2	Tellez *et al.*	**0.8170**	0.8121	0.8005	0.8099
3	Nieuwenhuis & Wilkens	0.7830	0.8116	0.8027	0.7991
4	Sierra-Loaiza & González	0.8120	0.8011	0.7827	0.7986
5	Ciccone *et al.*	0.7910	0.8074	0.7959	0.7981
6	Kosse *et al.*	0.7920	0.8074	0.7918	0.7971
7	Takahashi *et al.*	0.7710	0.7968	0.7864	0.7847
8	Veenhoven *et al.*	0.7490	0.7926	0.8036	0.7817
9	Martinc *et al*	0.7760	0.7900	0.7782	0.7814
10	López-Santillán *et al.*	0.7760	0.7847	0.7677	0.7761
11	Hacohen-Kerner *et al.* (B)	0.7590	0.7911	0.7650	0.7717
12	Hacohen-Kerner *et al.* (A)	0.7590	0.7911	0.7650	0.7717
13	Stout *et al.*	0.7600	0.7853	0.7405	0.7619
14	Gopal-Patra *et al.*	0.7430	0.7558	0.7586	0.7525
15	von Däniken *et al.*	0.7320	0.7742	0.7464	0.7509
16	Schaetti	0.7390	0.7711	0.7359	0.7487
	baseline-bow	0.7480	0.7411	0.7255	0.7382
17	Aragon & Lopez	0.6480	0.7963	0.7686	0.7376
18	Bayot & Gonçalves	0.6760	0.7716	0.6873	0.7116
19	Garibo	0.6750	0.7363	0.7164	0.7092
20	Sezerer *et al.*	0.6920	0.7495	0.6655	0.7023
21	Raiyani *et al.*	0.7220	0.7279	0.6436	0.6978
22	Sandroni-Dias & Paraboni	0.6870	0.6658	0.6782	0.6770
	baseline-stats	0.5000	0.5000	0.5000	0.5000
23	Karlgren *et al.*	-	0.5521	-	-
	Min	0.6480	0.5521	0.6436	0.6770
	Q1	0.7245	0.7634	0.7370	0.7404
	Median	0.7590	0.7900	0.7663	0.7717
	Mean	0.7485	0.7693	0.7546	0.7608
	SDev	0.0480	0.0586	0.0487	0.0399
	Q3	0.7812	0.7990	0.7904	0.7940
	Max	0.8170	0.8221	0.8200	0.8170
	Skewness	-0.5191	-2.5275	-0.8785	-0.5855
	Kurtosis	2.2985	9.5425	2.7640	2.2513
	Normality (p-value)	0.4126	0.0006	0.0757	0.1942

(74.80%) is very close to the mean (74.85%), while 9 participants are below. For English and Spanish, most participants were better than the baseline. For English, the obtained result (74.11%) is lower than the mean (76.93%) and even lower than the Q1 (76.34%), with 4 participants below (including the aforementioned outliers [26, 51]). For Spanish, the obtained result (72.55%) is below the mean (75.46%) and the Q1 (73.70%), with 5 participants below (including one outlier).

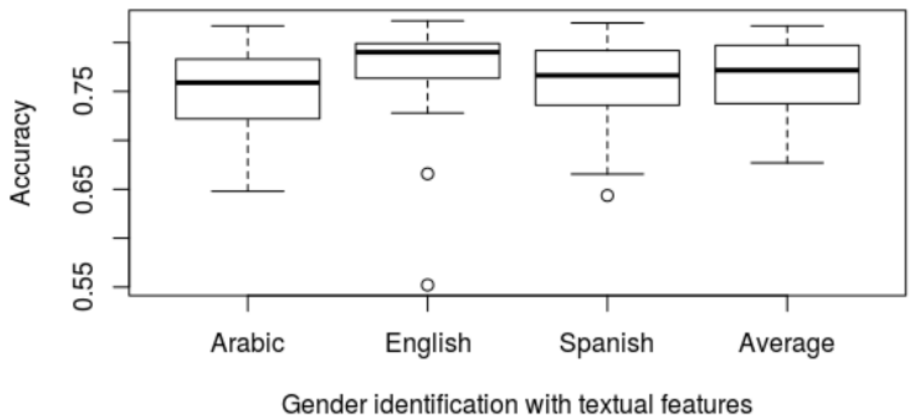

Figure 1. Distribution of the results for gender identification in the different languages when using text features only.

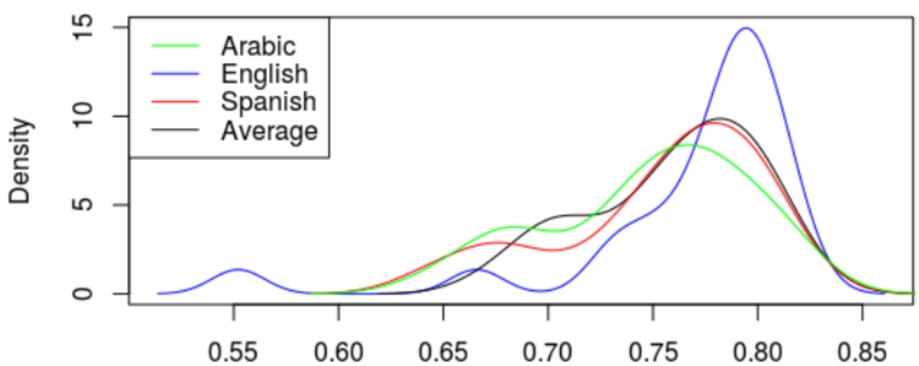

Figure 2. Density of the results for the gender identification in the different languages.

5.2 Gender Identification with Images

As can be seen in Table 3, the best results were achieved for English (81.63%), with statistical significance over Spanish (77.32%) and Arabic (77.80%). All best results stem from the authors in [60], who used a pre-trained CNN on the basis of ImageNet. Despite this higher value for the best obtained result for English, the distributions of accuracies are very similar for the three languages, as can be seen in the Figures 3 and 4. The mean values are of 62.37%, 63.41%, and 61.86% for Arabic, English, and Spanish respectively, with standard deviations below 10% and following a normal distribution.

For Arabic, the second best result (72.80%) has been obtained by the authors in [56], who used VGG16 and ResNet50 from ImageNet. The third best result (70.10%) has been obtained by the authors in [14]. Besides color histograms they have detected faces, objects, and local binary patterns. Although there is no statistical significance between them at 95% of confidence, there is with respect to the best result (not at 99%). For English, the second (74.42%) and third (69.63%) best results are from the authors in [56] and [14] respectively. In both cases the difference is statistically significant. Similarly, for Spanish the second (71%) and third (68.05%) best results are from the authors in [56] and [14] respectively. Again, the difference is statistically significant.

As before, we can discard the statistical baseline. Similarly, most of the participants have achieved better results than the RGB baseline (52.60% on average); two participants achieved slightly lower results (50.23% and 50.22%) [23]). For all languages the baseline (54.10%, 51.79%, and 51.91%) is below the respective Q1s (55.57%, 56.89%, and 56.40%). Also note that this baseline is only slightly better than the statistical one, we shows that it is not suitable for the task.

Table 3. Accuracy per language in the gender identification task with images.

Ranking	Team	Arabic	English	Spanish	Average
1	Takahashi *et al.*	**0.7720**	**0.8163**	**0.7732**	**0.7872**
2	Sierra-Loaiza & González	0.7280	0.7442	0.7100	0.7274
3	Ciccone *et al.*	0.7010	0.6963	0.6805	0.6926
4	Aragon & Lopez	0.6800	0.6921	0.6668	0.6796
5	Gopal-Patra *et al.*	0.6570	0.6747	0.6918	0.6745
6	Stout *et al.*	0.6230	0.6584	0.6232	0.6349
7	Nieuwenhuis & Wilkens	0.6230	0.6100	0.5873	0.6068
8	Tellez *et al.*	0.5900	0.5468	0.5691	0.5686
9	Schaetti	0.5430	0.5763	0.5782	0.5658
10	Martinc *et al.*	0.5600	0.5826	0.5486	0.5637
	baseline-rgb	0.5410	0.5179	0.5191	0.5260
11	Hacohen-Kerner *et al.* (B)	0.5100	0.4942	0.5027	0.5023
12	Hacohen-Kerner *et al.* (A)	0.4970	0.5174	0.4923	0.5022
	baseline-stats	0.5000	0.5000	0.5000	0.5000
	Min	0.4970	0.4942	0.4923	0.5022
	Q1	0.5557	0.5689	0.5640	0.5653
	Median	0.6230	0.6342	0.6052	0.6209
	Mean	0.6237	0.6341	0.6186	0.6255
	SDev	0.0873	0.0964	0.0869	0.0893
	Q3	0.6853	0.6932	0.6833	0.6828
	Max	0.7720	0.8163	0.7732	0.7872
	Skewness	0.1079	0.2716	0.1528	0.1984
	Kurtosis	1.9374	2.2109	2.0109	2.0636
	Normality (p-value)	0.9836	0.9031	0.7356	0.5964

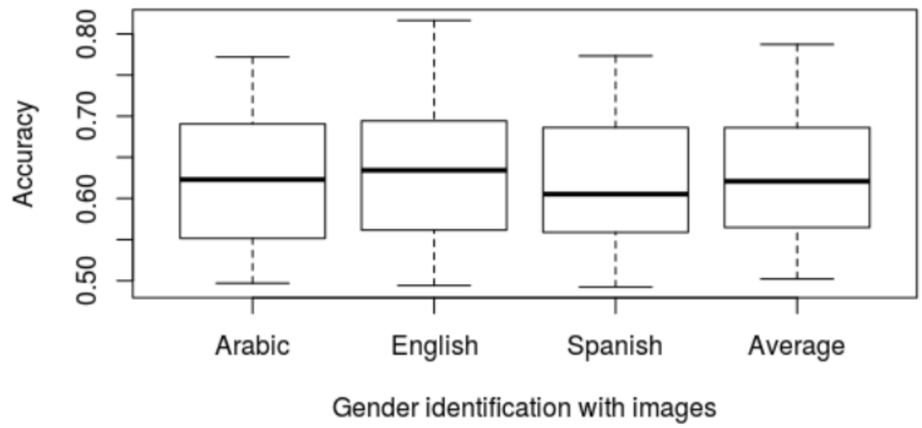

Figure 3. Distribution of the results for gender identification in the different languages when using images only.

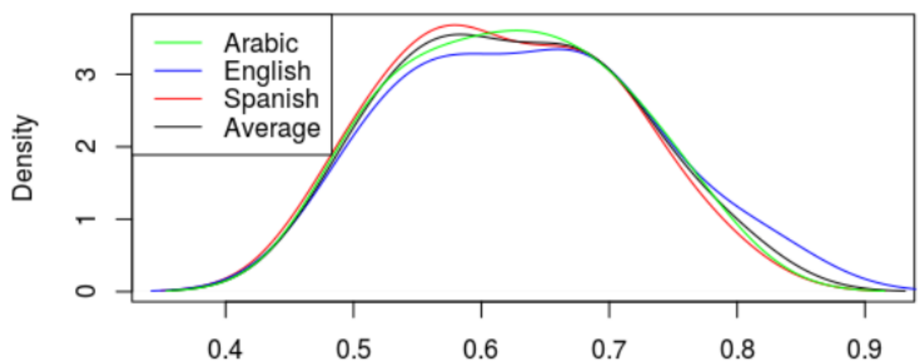

Figure 4. Density of the results for gender identification in the different languages.

5.3 Combined Approaches

We now analyse how images can help to tackle the gender identification task. Table 4 shows the basic statistics about the improvement (in %) for the different languages. On average, the improvement is very small (0.76% and 1.01% for Arabic and English), or even negative (-0.06%) for of Spanish. However, looking at Figure 5 it can be seen that some systems perform much better such as Takahashi *et al.*, who achieved an improvement of 7.73% for English.

Table 4. Distribution of the improvement over text classification in the different languages.

	Arabic	English	Spanish
Min	-0.2635	-0.6526	-4.4717
Q1	-0.0616	-0.0647	-0.6613
Median	0.3185	0.4249	0.0257
Mean	0.7613	1.0102	-0.0609
SDev	1.2513	2.2473	1.9087
Q3	0.8487	0.6788	0.4898
Max	3.3647	7.7309	3.7513
Skewness	1.2095	2.4716	-0.3778
Kurtosis	2.9616	8.0027	4.4883
Normality (p-value)	0.0010	0.0000	0.1316

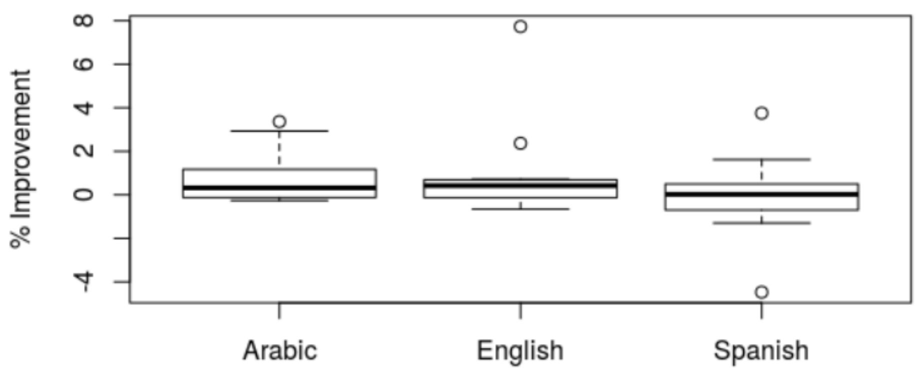

Figure 5. Distribution of the percentage of improvement over text classification.

The tables 5, 6, and 7 show the accuracies obtained with texts, with images, with their combination, and the percentage of improvement for Arabic, English, and Spanish respectively. Similarly, the Figures 6, 7, and 8 show for the same languages the density of the improvement distribution over text classification.

Table 5 shows the results for Arabic. As can be seen in Figure 6 the results do not follow a normal distribution; the improvement of most of the participants is between 0.53% and -0.26%, whereas three users obtain higher improvements: 1.82% [60], 2.93% [5], and 3.36% [38]. It is noteworthy that the systems that obtained the highest results tried to capture semantic features from images, and not only faces or colors. For example, Gopal-Patra *et al.* [38] used an image captioning system [38], Aragon & Lopez [5] ImageNet to obtain VGG16 features, and Takahashi *et al.* [60] a pre-trained CNN also on the basis of ImageNet.

Table 5. Improvement over text classification for Arabic.

Team	Texts	Images	Combined	Improvement
Gopal-Patra *et al.*	0.7430	0.6570	0.7680	3.3647%
Aragon & Lopez	0.6480	0.6800	0.6670	2.9321%
Takahashi *et al.*	0.7710	0.7720	0.7850	1.8158%
Stout *et al.*	0.7600	0.6230	0.7640	0.5263%
Nieuwenhuis & Wilkens	0.7830	0.6230	0.7870	0.5109%
Ciccone *et al.*	0.7910	0.7010	0.7940	0.3793%
Martinc *et al*	0.7760	0.5600	0.7780	0.2577%
Tellez *et al.*	0.8170	0.5900	0.8180	0.1224%
Schaetti	0.7390	0.5430	0.7390	0.0000%
Sierra-Loaiza & González	0.8120	0.7280	0.8100	-0.2463%
Hacohen-Kerner *et al.* (B)	0.7590	0.5100	0.7570	-0.2635%
Hacohen-Kerner *et al.* (A)	0.7590	0.4970	0.7570	-0.2635%

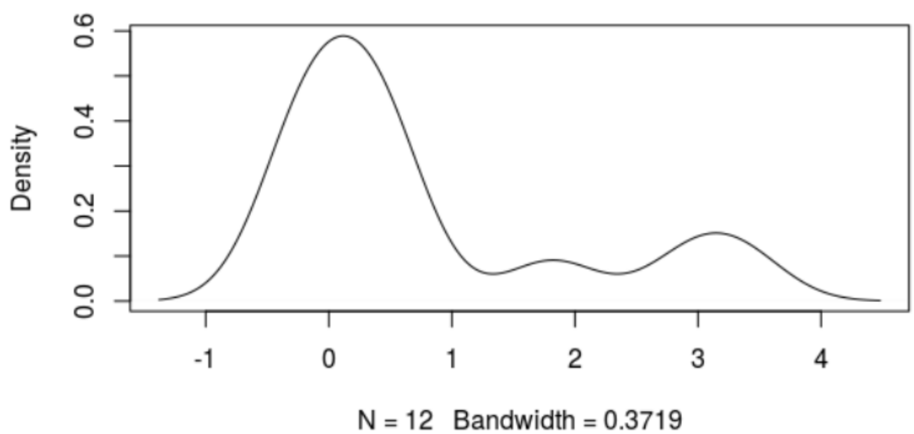

Figure 6. Density of the distribution of improvement over text classification for Arabic.

The distribution of improvements for English is even less normal, as can be seen in Figure 7. There are three groups of systems (see Table 6): *i)* systems with improvements between 0.72% and deteriorations of -4.65%, *ii)* one system with an improvement of 2.37% [38], and *iii)* one system with an improvement of 7.73% [60]. Similar to Arabic, the best results have been achieved by systems that exploit semantic features [60, 38]. Furthermore, the less negative results have been achieved either with the use of ImageNet and VGG16 features [5] or with the combination of face recognition, object detection, local binary patterns, and color histograms [14].

Table 6. Improvement over text classification for English.

Team	Texts	Images	Combined	Improvement
Takahashi *et al.*	0.7968	0.8163	0.8584	7.7309
Gopal-Patra *et al.*	0.7558	0.6747	0.7737	2.3684
Ciccone *et al.*	0.8074	0.6963	0.8132	0.7184
Aragon & Lopez	0.7963	0.6921	0.8016	0.6656
Sierra-Loaiza & González	0.8011	0.7442	0.8063	0.6491
Hacohen-Kerner *et al.* (A)	0.7911	0.5174	0.7947	0.4551
Stout *et al.*	0.7853	0.6584	0.7884	0.3948
Martinc *et al.*	0.7900	0.5826	0.7926	0.3291
Schaetti	0.7711	0.5763	0.7711	0.0000
Nieuwenhuis & Wilkens	0.8116	0.6100	0.8095	-0.2587
Hacohen-Kerner *et al.* (B)	0.7911	0.4942	0.7889	-0.2781
Tellez *et al.*	0.8121	0.5468	0.8068	-0.6526

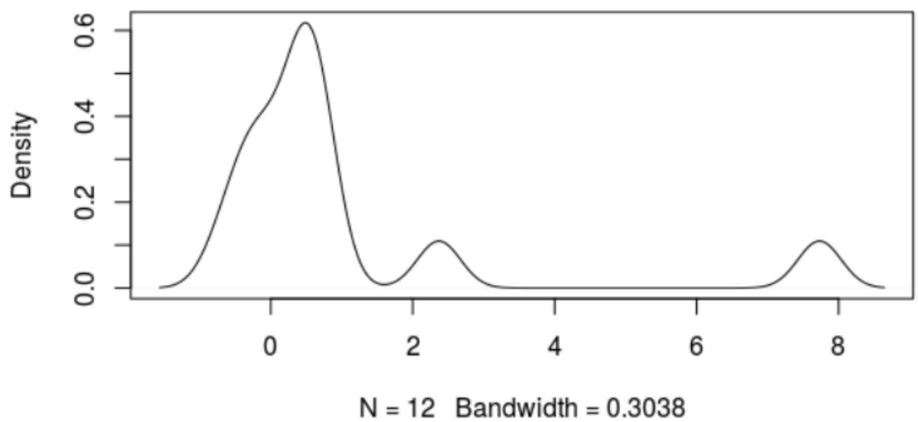

Figure 7. Density of the distribution of improvement over text classification for English.

For Spanish the systems' improvements follows a normal distribution, having two spikes in both extremes. In particular, there is *i)* one system whose deterioration is -4.47% [56], *ii)* a group of users with improvement/deterioration between -1.30% and 1.62%, and *iii)* one system with 3.75% of improvement [60]. In this regard, the best result has been obtained by Takahashi *et al.* with a pre-trained CNN from ImageNet, followed by the use of an image captioning system [38], the combination of faces, objects, and local binary patterns with color histograms [14], and the use of ImageNet to obtain VGG16 features [5].

Table 7. Improvement over text classification for Spanish.

Team	Texts	Images	Combined	Improvement
Takahashi *et al.*	0.7864	0.7732	0.8159	3.7513
Gopal-Patra *et al.*	0.7586	0.6918	0.7709	1.6214
Ciccone *et al.*	0.7959	0.6805	0.8000	0.5151
Aragon & Lopez	0.7686	0.6668	0.7723	0.4814
Stout *et al.*	0.7405	0.6232	0.7432	0.3646
Martinc *et al.*	0.7782	0.5486	0.7786	0.0514
Schaetti	0.7359	0.5782	0.7359	0.0000
Hacohen-Kerner *et al.* (A)	0.7650	0.4923	0.7623	-0.3529
Tellez *et al.*	0.8005	0.5691	0.7955	-0.6246
Hacohen-Kerner *et al.* (B)	0.7650	0.5027	0.7591	-0.7712
Nieuwenhuis & Wilkens	0.8027	0.5873	0.7923	-1.2956
Sierra-Loaiza & González	0.7827	0.7100	0.7477	-4.4717

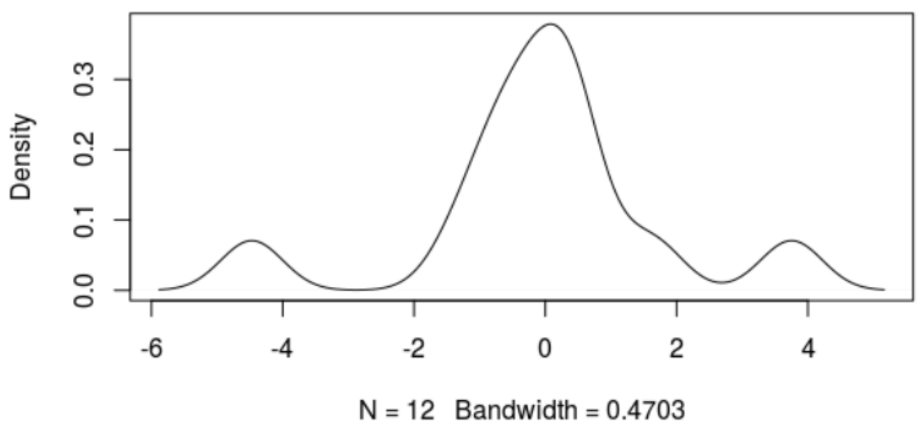

Figure 8. Density of the distribution of improvement over text classification for Spanish.

5.4 Final Ranking and Best Results

This year 23 teams participated in the shared task; Table 8 shows the overall performance per language and user's ranking. The best results have been obtained for English (85.84%), followed by Spanish (82%), and Arabic (81.80%).

Table 8. Accuracy per language and global ranking as average per language.

Ranking	Team	Arabic	English	Spanish	Average
1	Takahashi *et al.*	0.7850	**0.8584**	0.8159	**0.8198**
2	Daneshvar	0.8090	0.8221	**0.8200**	0.8170
3	Tellez *et al.*	**0.8180**	0.8068	0.7955	0.8068
4	Ciccone *et al.*	0.7940	0.8132	0.8000	0.8024
5	Kosse *et al.*	0.7920	0.8074	0.7918	0.7971
6	Nieuwenhuis & Wilkens	0.7870	0.8095	0.7923	0.7963
7	Sierra-Loaiza & González	0.8100	0.8063	0.7477	0.7880
8	Martinc *et al.*	0.7780	0.7926	0.7786	0.7831
9	Veenhoven *et al.*	0.7490	0.7926	0.8036	0.7817
10	López-Santillán *et al.*	0.7760	0.7847	0.7677	0.7761
11	Hacohen-Kerner *et al.* (A)	0.7570	0.7947	0.7623	0.7713
12	Gopal-Patra *et al.*	0.7680	0.7737	0.7709	0.7709
13	Hacohen-Kerner *et al.* (B)	0.7570	0.7889	0.7591	0.7683
14	Stout *et al.*	0.7640	0.7884	0.7432	0.7652
15	von Däniken *et al.*	0.7320	0.7742	0.7464	0.7509
16	Schaetti	0.7390	0.7711	0.7359	0.7487
17	Aragon & Lopez	0.6670	0.8016	0.7723	0.7470
18	Bayot & Gonçalves	0.6760	0.7716	0.6873	0.7116
19	Garibo	0.6750	0.7363	0.7164	0.7092
20	Sezerer *et al.*	0.6920	0.7495	0.6655	0.7023
21	Raiyani *et al.*	0.7220	0.7279	0.6436	0.6978
22	Sandroni-Dias & Paraboni	0.6870	0.6658	0.6782	0.6770
23	Karlgren *et al.*	-	0.5521	-	-
	Min	0.6670	0.5521	0.6436	0.6770
	Q1	0.7245	0.7713	0.7377	0.7474
	Median	0.7605	0.7889	0.7650	0.7711
	Mean	0.7515	0.7735	0.7543	0.7631
	SDev	0.0471	0.0614	0.0493	0.0409
	Q3	0.7865	0.8065	0.7922	0.7942
	Max	0.8180	0.8584	0.8200	0.8198
	Skewness	-0.4908	-2.2563	-0.7807	-0.6090
	Kurtosis	2.0346	8.7093	2.6912	2.3341
	Normality (p-value)	0.3490	0.0002	0.3341	0.1717

The overall best result (81.98%) is from the authors in [60] who approached the task with deep neural networks. For text processing, they used word embeddings from a stream of tweets with FastText skip-grams and trained a Recurrent Neural Network. For images, they used a pre-trained Convolutional Neural Network. They combined both approaches with a fusion component. The authors in [15] got the second best result on average (81.70%) by approaching the task only from the textual perspective. They used an SVM with different types of word and character n-grams. The third best overall result (80.68%) stems from the authors in [61]. They used an SVM with combinations of word and character n-grams for texts and a variant of the Bag of Visual Words for images, combining both predictions with a convex linear combination. According to t-Student, there is no statistical significance among the three approaches. This is also supported by the Bayesian Signed-Rank test [12] between Takahashi *et al.* and Daneshvar, as shown in Figure 9. However, for Takahashi *et al.* and Tellez *et al.*, the probability of the first system to perform better (62.96%) is higher than the sum of being equal (20.64%) or worse (16.39%), as shown in Figure 10. The complete results of this test are presented in the Appendix B.

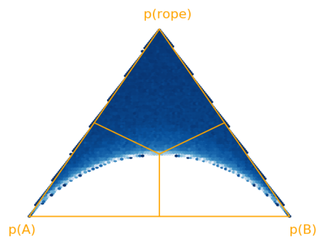

Figure 9. Bayesian Signed-Rank Test between Takahashi *et al.* and Daneshvar. P(A>B)=0.3416; P(A=B)=0.3191; P(A<B)=0.3392

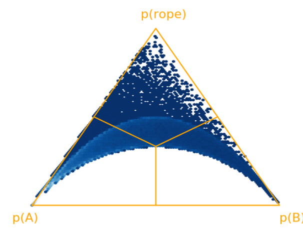

Figure 10. Bayesian Signed-Rank Test between Takahashi *et al.* and Tellez *et al.*. P(A>B)=0.6296; P(A=B)=0.2064; P(A<B)=0.1639

With respect to the different languages, the best results have been obtained by the same authors. The best results for Arabic (81.80%) stem from the authors in [61], the best results for English (85.84%) from the authors in [60], and the best results for Spanish (82%) from the authors in [15]. Note that the only result that is significantly higher is the one obtained for English (85.84%).

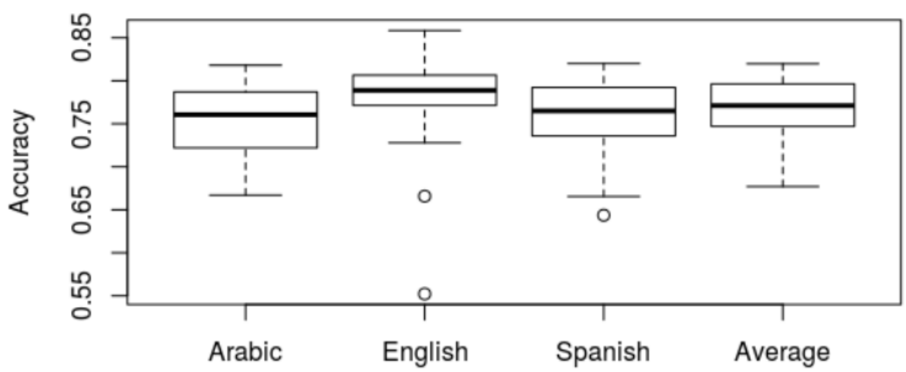

Figure 11. Distribution of the results for gender identification in the different languages.

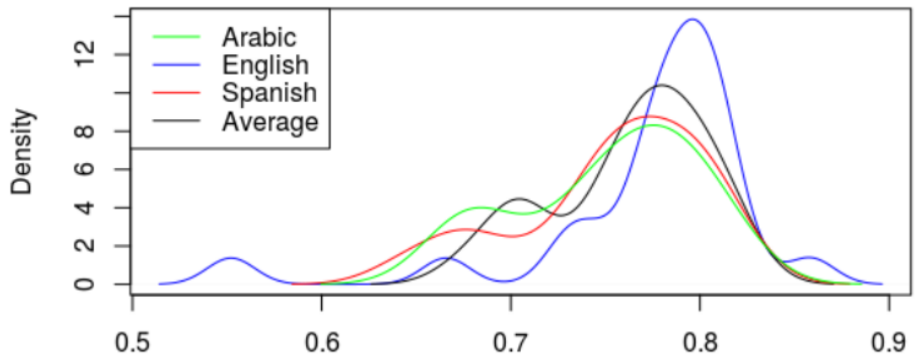

Figure 12. Distribution of the results for gender identification in the different languages.

Table 9 shows the best results per language and modality. The results achieved with the textual approach are higher than the results obtained with images, although being very similar to those for English. It should be highlighted that the best results were obtained by combining texts and images, where in the case of English the improvement is higher.

Table 9. Best results per language and modality.

Language	Textual	Images	Combined
Arabic	0.8170	0.7720	0.8180
English	0.8221	0.8163	0.8584
Spanish	0.8200	0.7732	0.8200

6 Conclusion

In this paper we presented the results of the 6th International Author Profiling Shared Task at PAN 2018, hosted at CLEF 2018. The participants had to identify the gender from Twitter authors, considering both a multimodal and a multilingual perspective: the provided data contains both tweets and images and cover the three languages Arabic, English, and Spanish.

The participants used different approaches to tackle the task, with deep learning approaches prevailing. However, the best results regarding the textual subtask have been obtained with combinations of different types of n-grams and traditional machine learning algorithms such as SVM and Logistic Regression. Only the second best result for Spanish was obtained with a bi-LSTM, which has been trained with word embeddings.

For the classification of images the approaches can be grouped in three types: *i)* approaches based on face recognition, *ii)* approaches based on pre-trained models and image processing tools such as ImageNet, and *iii)* approaches with "hand-crafted" features such as color histograms and bag-of-visual-words. Regarding the second type, the best results were obtained with semantic features extracted from the images. Approaches based on face recognition do not belong to the best, which may be rooted in the fact that many images do not show faces—and if, the contained faces do not depict the author.

According to the achieved results, text features discriminate better between genders than do images. However, the combined use of both modalities provides insights: On average, there is no improvement when images are used, which is due to the low performance of some inferior approaches. However, for more elaborated representations, which obtain semantics from the images with the use of tools such as ImageNet, the improvement is up to 7.73% for English (taking into account that the accuracy obtained only with text features is even high).

The best results in the shared tasks are over 80% on average, with the highest result for English (85.84%) [60], followed by Spanish (82%) [15], and Arabic (81.80%) [61]. Takahashi *et al.* [60] approached the task with deep learning techniques: word embeddings and RNN for texts and ImageNet-based CNN for images. Daneshvar [15] approached the task using the textual modality only. The author trained an SVM with combinations of word and character n-grams. Finally, Tellez *et al.* [61] used SVM with different kinds of n-grams, combined with a variant of the Bag of Visual Words (BoVW) using the DAISY feature descriptor. Altogether, traditional approaches still remain competitive, while some new approaches based on deep learning are acquiring strength.

Acknowledgements

Our special thanks goes to all PAN participants for providing high-quality submission, and to MeaningCloud[13] for sponsoring the author profiling shared task award. The first author acknowledges the SomEMBED TIN2015-71147-C2-1-P MINECO research project. The third author acknowledges the CONACyT FC-2016/2410. The work on the data for Arabic as well as this publication were made possible by NPRP grant #9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). Responsible for the statements made herein are the first two authors.

Bibliography

[1] Kholoud Alsmearat, Mahmoud Al-Ayyoub, and Riyad Al-Shalabi. An extensive study of the bag-of-words approach for gender identification of arabic articles. In *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, pages 601–608. IEEE, 2014.

[2] Kholoud Alsmearat, Mohammed Shehab, Mahmoud Al-Ayyoub, Riyad Al-Shalabi, and Ghassan Kanaan. Emotion analysis of arabic articles and its impact on identifying the author's gender. In *Computer Systems and Applications (AICCSA), 2015 IEEE/ACS 12th International Conference on*, 2015.

[3] Emad AlSukhni and Qasem Alequr. Investigating the use of machine learning algorithms in detecting gender of the arabic tweet author. *International Journal of Advanced Computer Science & Applications*, 1(7):319–328, 2016.

[13] http://www.meaningcloud.com/

[4] Miguel-Angel Álvarez-Carmona, A.-Pastor López-Monroy, Manuel Montes-Y-Gómez, Luis Villaseñor-Pineda, and Hugo Jair-Escalante. Inaoe's participation at pan'15: author profiling task—notebook for pan at clef 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Labs and Workshops, Notebook Papers*, CEUR Workshop Proceedings. CEUR-WS.org, 2015. URL http://www.clef-initiative.eu/publication/working-notes.

[5] Mario Ezra Aragón and A.-Pastor López-Monroy. A straightforward multimodal approach for author profiling. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[6] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *TEXT*, 23:321–346, 2003.

[7] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*, 2017.

[8] Samiul Azam and Marina Gavrilova. Gender prediction using individual perceptual image aesthetics. 2016.

[9] Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. N-gram: New groningen author-profiling model. *arXiv preprint arXiv:1707.03764*, 2017.

[10] Roy Bayot and Teresa Gonçalves. Multilingual author profiling using word embedding averages and svms. In *Software, Knowledge, Information Management & Applications (SKIMA), 2016 10th International Conference on*, pages 382–386. IEEE, 2016.

[11] Roy Khristopher Bayot and Teresa Gon calves. Multilingual author profiling using lstms. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[12] A. Benavoli, F. Mangili, G. Corani, M. Zaffalon, and F. Ruggeri. A Bayesian Wilcoxon signed-rank test based on the Dirichlet process. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2014)*, pages 1–9, 2014. URL http://www.idsia.ch/ alessio/benavoli2014a.pdf.

[13] John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[14] Giovanni Ciccone, Arthur Sultan, Léa Laporte, Elöd Egyed-Zsigmond, Alaa Alhamzeh, and Michael Granitzer. Stacked gender prediction from tweet texts and images. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro,

editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[15] Saman Daneshvar. Gender identification in twitter using n-grams and lsa. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[16] Makinen Erno, Roope Raisamo, et al. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):541–547, 2007.

[17] Dominique Estival, Tanja Gaustad, Ben Hutchinson, Son Bao Pham, and Will Radford. Author profiling for english and arabic emails. 2008.

[18] Francisco Manuel, Rangel Pardo, Paolo Rosso, Martin Potthast, and Benno Stein. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomad Mandl, editors, *Working Notes Papers of the CLEF 2017 Evaluation Labs*, volume 1866 of *CEUR Workshop Proceedings*. CLEF and CEUR-WS.org, September 2017. URL http://ceur-ws.org/Vol-1866/.

[19] Òscar Garibo-Orts. A big data approach to gender classification in twitter. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[20] Tim Gollub, Benno Stein, and Steven Burrows. Ousting ivory tower research: towards a web framework for providing experiments as a service. In Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, pages 1125–1126. ACM, August 2012. ISBN 978-1-4503-1472-5.

[21] Tim Gollub, Benno Stein, Steven Burrows, and Dennis Hoppe. TIRA: Configuring, executing, and disseminating information retrieval experiments. In A Min Tjoa, Stephen Liddle, Klaus-Dieter Schewe, and Xiaofang Zhou, editors, *9th International Workshop on Text-based Information Retrieval (TIR 12) at DEXA*, pages 151–155, Los Alamitos, California, September 2012. IEEE. ISBN 978-1-4673-2621-6.

[22] Tim Gollub, Martin Potthast, Anna Beyer, Matthias Busse, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Recent trends in digital text forensics and its evaluation. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative (CLEF 13)*, pages 282–302, Berlin Heidelberg New York, September 2013. Springer. ISBN 978-3-642-40801-4.

[23] Yaakov HaCohen-Kerner, Yair Yigal, Elyashiv Shayovitz, Daniel Miller 1, and Toby Breckon. Author profiling: Gender prediction from tweets and images. In

Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Janet Holmes and Miriam Meyerhoff. *The handbook of language and gender*. Blackwell Handbooks in Linguistics. Wiley, 2003.

[26] Jussi Karlgren, Lewis Esposito, Chantal Gratton, and Pentti Kanerva. Authorship profiling without using topical information. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[27] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. literary and linguistic computing 17(4), 2002.

[28] Rick Kosse, Youri Schuur, and Guido Cnossen. Mixing traditional methods with neural networks for gender prediction. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[29] A. Pastor Lopez-Monroy, Manuel Montes-Y-Gomez, Hugo Jair Escalante, Luis Villasenor-Pineda, and Esau Villatoro-Tello. INAOE's participation at PAN'13: author profiling task—Notebook for PAN at CLEF 2013. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain*, September 2013.

[30] A. Pastor López-Monroy, Manuel Montes y Gómez, Hugo Jair-Escalante, and Luis Villase nor Pineda. Using intra-profile information for author profiling—Notebook for PAN at CLEF 2014. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*, September 2014.

[31] A. Pastor López-Monroy, Manuel Montes y Gómez, Hugo Jair-Escalante, Luis Villase nor Pineda, and Thamar Solorio. Uh-inaoe participation at pan17: Author profiling. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of the CLEF 2017 Evaluation Labs*. CLEF and CEUR-WS.org, September 2017.

[32] Roberto López-Santillán, Luis-Carlos González-Gurrola, and Graciela Ramírez-Alonso. Custom document embeddings via the centroids method: Gender classification in an author profiling task. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets*

Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), September 2018.

[33] Xiaojun Ma, Yukihiro Tsuboshita, and Noriji Kato. Gender estimation for sns user profiling using automatic image annotation. In *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, pages 1–6. IEEE, 2014.

[34] Suraj Maharjan, Prasha Shrestha, Thamar Solorio, and Ragib Hasan. A straightforward author profiling approach in mapreduce. In *Advances in Artificial Intelligence. Iberamia*, pages 95–107, 2014.

[35] Matej Martinc, Blaẑ Ŝkrlj, and Senja Pollak. Multilingual gender classification with multi-view deep learning. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[36] Baback Moghaddam and Ming-Hsuan Yang. Gender classification with support vector machines. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 306–311. IEEE, 2000.

[37] Moniek Nieuwenhuis and Jeroen Wilkens. Twitter text and image gender classification with a logistic regression n-gram model. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[38] Braja Gopal Patra, Kumar Gourav Das, and Dipankar Das. Multimodal author profiling for arabic, english, and spanish. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[39] James W. Pennebaker. *The secret life of pronouns: what our words say about us*. Bloomsbury USA, 2013.

[40] James W. Pennebaker, Mathias R. Mehl, and Kate G. Niederhoffer. Psychological aspects of natural language use: our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[41] Kashyap Raiyani, Paulo Quaresma Teresa Gončalves, and Vitor Beires-Nogueira. Multi-language neural network model with advance preprocessor for gender classification over social media. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[42] Francisco Rangel and Paolo Rosso. On the multilingual and genre robustness of emographs for author profiling in social media. In *6th international conference of CLEF on experimental IR meets multilinguality, multimodality, and interaction*, pages 274–280. Springer-Verlag, LNCS(9283), 2015.

[43] Francisco Rangel and Paolo Rosso. On the impact of emotions on author profiling. *Information processing & management*, 52(1):73–92, 2016.

[44] Francisco Rangel, Paolo Rosso, Moshe Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at pan 2013. In *Forner P., Navigli R., Tufis D. (Eds.), CLEF 2013 labs and workshops, notebook papers. CEUR-WS.org, vol. 1179*, 2013.

[45] Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the 2nd author profiling task at pan 2014. In *Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 labs and workshops, notebook papers. CEUR-WS.org, vol. 1180*, 2014.

[46] Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd author profiling task at pan 2015. In *Cappellato L., Ferro N., Jones G., San Juan E. (Eds.) CLEF 2015 labs and workshops, notebook papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1391*, 2015.

[47] Francisco Rangel, Paolo Rosso, and Marc Franco-Salvador. A low dimensionality representation for language variety identification. In *17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing*. Springer-Verlag, LNCS, arXiv:1705.10754, 2016.

[48] Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 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.) CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1866.*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2016.

[49] Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2016.

[50] Charles A Russell and Bowman H Miller. Profile of a terrorist. *Studies in Conflict & Terrorism*, 1(1):17–34, 1977.

[51] Rafael-Felipe Sandroni-Dias and Ivandré Paraboni. Author profiling using word embeddings with subword information. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[52] Nils Schaetti. Unine at clef 2018: Character-based convolutional neural network and resnet18 for twitter author profiling. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality,*

Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), September 2018.

[53] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205. AAAI, 2006.

[54] Erhan Sezerer, Ozan Polatbilek, Özge Sevgili, and Selma Tekir. Gender prediction from tweets with convolutional neural networks. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[55] Ryosuke Shigenaka, Yukihiro Tsuboshita, and Noriji Kato. Content-aware multi-task neural networks for user gender inference based on social media images. In *Multimedia (ISM), 2016 IEEE International Symposium on*, pages 169–172. IEEE, 2016.

[56] Sebastián Sierra-Loaiza and Fabio A. González. Combining textual and visual representations for multimodal author profiling. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[58] Luka Stout, Robert Musters, and Chris Pool. Author profiling based on text and images. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[59] Zehang Sun, George Bebis, Xiaojing Yuan, and Sushil J Louis. Genetic feature subset selection for gender classification: A comparison study. In *Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on*, pages 165–170. IEEE, 2002.

[60] Takumi Takahashi, Takuji Tahara, Koki Nagatani, Yasuhide Miura, Tomoki Taniguchi, and Tomoko Ohkuma. Text and image synergy with feature cross technique for gender identification. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[61] Eric S. Tellez, Sabino Miranda-Jiménez, Daniela Moctezuma, Mario Graff, Vladimir Salgado, and José Ortiz-Bejar. Gender identification through multi-modal tweet analysis using microtc and bag of visual words. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure

Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[62] Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):815–830, 2010.

[63] Satoshi Tsutsui and David Crandall. Using artificial tokens to control languages for multilingual image caption generation. *arXiv preprint arXiv:1706.06275*, 2017.

[64] Robert Veenhoven, Stan Snijders, Daniël van der Hall, and Rik van Noord. Using translated data to improve deep learning author profiling models. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[65] Pius von Däniken, Ralf Grubenmann, and Mark Cieliebak. Word unigram weighing for author profiling at pan 2018. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, September 2018.

[66] Edson Weren, Anderson Kauer, Lucas Mizusaki, Viviane Moreira, Palazzo de Oliveira, and Leandro Wives. Examining multiple features for author profiling. In *Journal of Information and Data Management*, pages 266–279, 2014.

Appendix A Pairwise Comparison of all Systems

For all subsequent tables, the significance levels are encoded as follows:

Symbol	Significance Level		
-		∼	no evaluated
=	p > 0.05	∼	not significant
*	0.05 ≥ p > 0.01	∼	significant
**	0.01 ≥ p > 0.001	∼	very significant
***	p ≤ 0.001	∼	highly significant

	Aragon	Bayot	Ciccone	Daneshvar	Garibo	Gopal	Hacohen-Kerner (A)	Hacohen-Kerner (B)	Kosse	Lopez-Santillan	Martinc	Nieuwenhuis	Raiyani	Sandroni-Dias	Schaetti	Sezerer	Sierra-Loaiza	Stout	Takahashi	Tellez	Veenhoven	Von-Daniken
Aragon		=	***	***	=	***	***	***	***	***	***	***	***	*	***	*	***	***	***	***	***	***
Bayot			***	***	=	***	***	***	***	***	***	***	**	=	***	=	***	***	***	***	***	***
Ciccone				=	***	***	*	*	=	=	=	=	***	***	***	***	=	*	=	*	**	***
Daneshvar					***	***	***	***	=	**	**	**	***	***	***	***	=	***	***	=	***	***
Garibo						***	***	***	***	***	***	***	*	=	***	=	***	***	***	***	***	**
Gopal							=	=	***	*	*	**	=	**	=	**	***	=	*	***	=	=
Hacohen-Kerner (A)								=	**	=	=	=	*	***	=	***	***	=	=	***	=	=
Hacohen-Kerner (B)									**	=	=	=	*	***	=	***	***	=	=	***	=	=
Kosse										=	=	=	***	***	***	***	=	*	=	*	**	***
Lopez-Santillan											=	=	***	***	*	***	**	=	=	**	=	**
Martinc												=	***	***	*	***	**	=	=	**	=	**
Nieuwenhuis													***	***	***	***	**	=	=	**	*	***
Raiyani														*	=	=	***	**	**	***	=	=
Sandroni-Dias															**	=	***	***	***	***	***	**
Schaetti																**	***	=	*	***	=	=
Sezerer																	***	***	***	***	**	*
Sierra-Loaiza																		***	**	=	***	***
Stout																			=	***	=	*
Takahashi																				***	=	**
Tellez																					***	***
Veenhoven																						=
Von-Daniken																						

Table A1. Significance of accuracy differences between system pairs. Textual modality in Arabic.

	Aragon	Ciccone	Gopal	Hacohen-Kerner (A)	Hacohen-Kerner (B)	Martinc	Nieuwenhuis	Schaetti	Sierra-Loaiza	Stout	Takahashi	Tellez
Aragon		=	=	***	***	***	**	***	**	**	***	***
Ciccone			*	***	***	***	***	***	=	***	***	***
Gopal				***	***	***	=	***	***	=	***	**
Hacohen-Kerner (A)					=	**	***	=	***	***	***	***
Hacohen-Kerner (B)						*	***	=	***	***	***	***
Martinc							**	=	***	**	***	=
Nieuwenhuis								***	***	=	***	=
Schaetti									***	***	***	*
Sierra-Loaiza										***	***	***
Stout											***	=
Takahashi												***
Tellez												

Table A2. Significance of accuracy differences between system pairs. Image modality in Arabic.

	Aragon	Ciccone	Gopal	Hacohen-Kerner (A)	Hacohen-Kerner (B)	Martinc	Nieuwenhuis	Schaetti	Sierra-Loaiza	Stout	Takahashi	Tellez
Aragon		***	***	***	***	***	***	***	***	***	***	***
Ciccone			=	**	**	=	=	***	=	*	=	=
Gopal				=	=	=	=	=	**	=	=	***
Hacohen-Kerner (A)					=	=	*	=	***	=	=	***
Hacohen-Kerner (B)						=	*	=	***	=	=	***
Martinc							=	**	*	=	=	**
Nieuwenhuis								***	=	=	=	*
Schaetti									***	=	**	***
Sierra-Loaiza										**	=	=
Stout											=	***
Takahashi												**
Tellez												

Table A3. Significance of accuracy differences between system pairs. Combined modality in Arabic.

	Aragon	Bayot	Ciccone	Daneshvar	Garibo	Gopal	Hacohen-Kerner (A)	Hacohen-Kerner (B)	Karlgren	Kosse	Lopez-Santillan	Martinc	Nieuwenhuis	Raiyani	Sandroni-Dias	Schaetti	Sezerer	Sierra-Loaiza	Stout	Takahashi	Tellez	Veenhoven	Von-Daniken
Aragon		**	=	**	***	***	=	=	***	=	=	=	=	***	***	*	***	=	=	=	=	=	*
Bayot			***	***	**	=	=	=	***	***	=	=	***	***	***	=	*	**	=	**	***	*	=
Ciccone				*	***	***	=	=	***	=	**	*	=	***	***	***	***	=	*	=	=	=	***
Daneshvar					***	***	***	***	***	=	***	***	=	***	***	***	***	*	***	**	=	***	***
Garibo						=	***	***	***	***	***	***	***	=	***	**	=	***	***	***	***	***	***
Gopal							**	**	***	***	**	**	***	**	***	=	=	***	**	***	***	***	=
Hacohen-Kerner (A)								=	***	=	=	=	*	***	***	=	***	=	=	=	*	=	=
Hacohen-Kerner (B)									***	=	=	=	*	***	***	*	***	=	=	=	*	=	=
Karlgren										***	***	***	***	***	***	***	***	***	***	***	***	***	***
Kosse											**	*	=	***	***	***	***	=	*	=	=	=	***
Lopez-Santillan												=	***	***	***	=	**	=	=	=	**	=	=
Martinc													*	***	***	=	***	=	=	=	*	=	=
Nieuwenhuis														***	***	***	***	=	**	=	=	*	***
Raiyani															***	***	=	***	***	***	***	***	***
Sandroni-Dias																***	***	***	***	***	***	***	***
Schaetti																	*	**	=	**	***	*	=
Sezerer																		***	***	***	***	***	*
Sierra-Loaiza																			=	=	=	=	**
Stout																				=	**	=	=
Takahashi																					=	=	*
Tellez																						*	***
Veenhoven																							=
Von-Daniken																							

Table A4. Significance of accuracy differences between system pairs. Textual modality in English.

	Aragon	Ciccone	Gopal	Hacohen-Kerner (A)	Hacohen-Kerner (B)	Martinc	Nieuwenhuis	Schaetti	Sierra-Loaiza	Stout	Takahashi	Tellez
Aragon		=	=	***	***	***	***	***	***	*	***	***
Ciccone			=	***	***	***	***	***	***	**	***	***
Gopal				***	***	***	***	***	***	=	***	***
Hacohen-Kerner (A)					=	***	***	***	***	***	***	=
Hacohen-Kerner (B)						***	***	***	***	***	***	**
Martinc							=	=	***	***	***	*
Nieuwenhuis								*	***	**	***	***
Schaetti									***	***	***	=
Sierra-Loaiza										***	***	***
Stout											***	***
Takahashi												***
Tellez												

Table A5. Significance of accuracy differences between system pairs. Image modality in English.

	Aragon	Ciccone	Gopal	Hacohen-Kerner (A)	Hacohen-Kerner (B)	Martinc	Nieuwenhuis	Schaetti	Sierra-Loaiza	Stout	Takahashi	Tellez
Aragon		=	**	=	=	=	=	**	=	=	***	=
Ciccone			***	*	**	*	=	***	=	**	***	=
Gopal				=	=	=	***	=	**	=	***	**
Hacohen-Kerner (A)					=	=	=	*	=	=	***	=
Hacohen-Kerner (B)						=	*	=	=	=	***	=
Martinc							=	*	=	=	***	=
Nieuwenhuis								***	=	*	***	=
Schaetti									**	=	***	***
Sierra-Loaiza										=	***	=
Stout											***	=
Takahashi												***
Tellez												

Table A6. Significance of accuracy differences between system pairs. Combined modality in English.

	Aragon	Bayot	Ciccone	Daneshvar	Garibo	Gopal	Hacohen-Kerner (A)	Hacohen-Kerner (B)	Kosse	Lopez-Santillan	Martinc	Nieuwenhuis	Raiyani	Sandroni-Dias	Schaetti	Sezerer	Sierra-Loaiza	Stout	Takahashi	Tellez	Veenhoven	Von-Daniken
Aragon		***	**	***	***	=	=	=	*	=	=	***	***	***	***	***	=	**	=	***	***	*
Bayot			***	***	**	***	***	***	***	***	***	***	***	=	***	=	***	***	***	***	***	***
Ciccone				***	***	***	***	***	=	**	*	=	***	***	***	***	=	***	=	=	=	***
Daneshvar					***	***	***	***	***	***	***	**	***	***	***	***	***	***	***	*	*	***
Garibo						***	***	***	***	***	***	***	***	**	=	***	***	*	***	***	***	**
Gopal							=	=	***	=	*	***	***	***	*	***	*	=	**	***	***	=
Hacohen-Kerner (A)								=	**	=	=	***	***	***	**	***	*	*	*	***	***	=
Hacohen-Kerner (B)									**	=	=	***	***	***	**	***	*	*	*	***	***	=
Kosse										**	=	=	***	***	***	***	=	***	=	=	=	***
Lopez-Santillan											=	***	***	***	**	***	=	**	*	***	***	*
Martinc												**	***	***	***	***	=	***	=	*	**	***
Nieuwenhuis													***	***	***	***	*	***	=	=	=	***
Raiyani														**	***	=	***	***	***	***	***	***
Sandroni-Dias															***	=	***	***	***	***	***	***
Schaetti																***	***	=	***	***	***	=
Sezerer																	***	***	***	***	***	***
Sierra-Loaiza																		***	=	=	*	***
Stout																			***	***	***	=
Takahashi																				=	=	***
Tellez																					=	***
Veenhoven																						***
Von-Daniken																						

Table A7. Significance of accuracy differences between system pairs. Textual modality in Spanish.

	Aragon	Ciccone	Gopal	Hacohen-Kerner (A)	Hacohen-Kerner (B)	Martinc	Nieuwenhuis	Schaetti	Sierra-Loaiza	Stout	Takahashi	Tellez
Aragon		=	*	***	***	***	***	***	***	***	***	***
Ciccone			=	***	***	***	***	***	**	***	***	***
Gopal				***	***	***	***	***	=	***	***	***
Hacohen-Kerner (A)					=	***	***	***	***	***	***	***
Hacohen-Kerner (B)						**	***	***	***	***	***	***
Martinc							**	***	***	***	***	=
Nieuwenhuis								=	***	*	***	=
Schaetti									***	***	***	=
Sierra-Loaiza										***	***	***
Stout											***	***
Takahashi												***
Tellez												

Table A8. Significance of accuracy differences between system pairs. Image modality in Spanish.

	Aragon	Ciccone	Gopal	Hacohen-Kerner (A)	Hacohen-Kerner (B)	Martinc	Nieuwenhuis	Schaetti	Sierra-Loaiza	Stout	Takahashi	Tellez
Aragon		**	=	=	=	=	*	***	*	**	***	*
Ciccone			**	***	***	**	=	***	***	***	=	=
Gopal				=	=	=	*	**	*	**	***	*
Hacohen-Kerner (A)					=	=	***	*	=	=	***	**
Hacohen-Kerner (B)						*	***	*	=	=	***	***
Martinc							=	***	**	**	***	=
Nieuwenhuis								***	***	***	*	=
Schaetti									=	=	***	***
Sierra-Loaiza										=	***	***
Stout											***	***
Takahashi												*
Tellez												

Table A9. Significance of accuracy differences between system pairs. Combined modality in Spanish.

Appendix B Bayesian Signed-Rank Test Among Systems

Team(A)	Team (B)	P(A>B)	P(A=B)	P(A<B)
Takahashi	Daneshvar	0.3416	0.3191	0.3392
Takahashi	Tellez	0.6296	0.2064	0.1639
Takahashi	Ciccone	0.5839	0.4161	0.0000
Takahashi	Kosse	0.6920	0.3079	0.0000
Takahashi	Nieuwenhuis	0.8435	0.1565	0.0000
Takahashi	Sierra-Loaiza	0.8702	0.0201	0.1096
Takahashi	Martinc	0.8414	0.1586	0.0000
Takahashi	Veenhoven	0.9533	0.0467	0.0000
Takahashi	López-Santillán	0.8423	0.1577	0.0000
Takahashi	Hacohen (A)	0.9886	0.0114	0.0000
Takahashi	Gopal-Patra	0.9518	0.0482	0.0000
Takahashi	Hacohen (B)	0.9882	0.0118	0.0000
Takahashi	Stout	0.9888	0.0112	0.0000
Takahashi	Von Däniken	0.9882	0.0118	0.0000
Takahashi	Schaetti	0.9886	0.0113	0.0000
Takahashi	Aragon	0.9878	0.0122	0.0000
Takahashi	Bayot	0.9883	0.0116	0.0000
Takahashi	Garibo	0.9888	0.0112	0.0000
Takahashi	Sezerer	0.9888	0.0112	0.0000
Takahashi	Raiyani	0.9881	0.0119	0.0000
Takahashi	Sandroni	0.9884	0.0116	0.0000
Daneshvar	Tellez	0.3999	0.6001	0.0000
Daneshvar	Ciccone	0.6535	0.3469	0.0000
Daneshvar	Kosse	0.8872	0.1128	0.0000
Daneshvar	Nieuwenhuis	0.9527	0.0473	0.0000
Daneshvar	Sierra-Loaiza	0.5857	0.4143	0.0000
Daneshvar	Martinc	0.9886	0.0114	0.0000
Daneshvar	Veenhoven	0.9522	0.0478	0.0000
Daneshvar	López-Santillán	0.9885	0.0115	0.0000
Daneshvar	Hacohen (A)	0.9883	0.01173	0.0000
Daneshvar	Gopal-Patra	0.9884	0.01159	0.0000
Daneshvar	Hacohen (B)	0.9886	0.01135	0.0000
Daneshvar	Stout	0.9884	0.01158	0.0000
Daneshvar	Von Däniken	0.9882	0.0118	0.0000
Daneshvar	Schaetti	0.9889	0.0111	0.0000
Daneshvar	Aragon	0.9883	0.0117	0.0000
Daneshvar	Bayot	0.9882	0.0118	0.0000
Daneshvar	Garibo	0.9883	0.0117	0.0000
Daneshvar	Sezerer	0.9886	0.0114	0.0000
Daneshvar	Raiyani	0.98854	0.0115	0.0000
Daneshvar	Sandroni	0.9884	0.0115	0.0000
Tellez	Ciccone	0.0943	0.9057	0.0000
Tellez	Kosse	0.4059	0.5941	0.0000
Tellez	Nieuwenhuis	0.4065	0.5935	0.0000
Tellez	Sierra-Loaiza	0.4065	0.5935	0.0000
Tellez	Martinc	0.8871	0.1129	0.0000
Tellez	Veenhoven	0.5862	0.4138	0.0000
Tellez	López-Santillán	0.9887	0.0113	0.0000
Tellez	Hacohen (A)	0.9535	0.0465	0.0000
Tellez	Gopal-Patra	0.9885	0.0115	0.0000
Tellez	Hacohen (B)	0.9527	0.0473	0.0000
Tellez	Stout	0.9522	0.0477	0.0000
Tellez	Von Däniken	0.9883	0.0117	0.0000
Tellez	Schaetti	0.9884	0.0116	0.0000
Tellez	Aragon	0.8446	0.1554	0.0000
Tellez	Bayot	0.9886	0.0114	0.0000
Tellez	Garibo	0.9883	0.0117	0.0000
Tellez	Sezerer	0.9884	0.0116	0.0000
Tellez	Raiyani	0.9883	0.0117	0.0000
Tellez	Sandroni	0.9882	0.0117	0.0000

Team(A)	Team (B)	P(A>B)	P(A=B)	P(A<B)
Ciccone	Kosse	0.0000	1.0000	0.0000
Ciccone	Nieuwenhuis	0.0000	1.0000	0.0000
Ciccone	Sierra-Loaiza	0.4580	0.4730	0.0690
Ciccone	Martinc	0.9531	0.0468	0.0000
Ciccone	Veenhoven	0.6941	0.3059	0.0000
Ciccone	López-Santillán	0.9543	0.0457	0.0000
Ciccone	Hacohen (A)	0.9530	0.0470	0.0000
Ciccone	Gopal-Patra	0.9887	0.0113	0.0000
Ciccone	Hacohen (B)	0.9886	0.0114	0.0000
Ciccone	Stout	0.9888	0.0112	0.0000
Ciccone	Von Däniken	0.9886	0.0114	0.0000
Ciccone	Schaetti	0.9883	0.0117	0.0000
Ciccone	Aragon	0.9530	0.0470	0.0000
Ciccone	Bayot	0.9883	0.0117	0.0000
Ciccone	Garibo	0.9884	0.0116	0.0000
Ciccone	Sezerer	0.9879	0.0121	0.0000
Ciccone	Raiyani	0.9885	0.0115	0.0000
Ciccone	Sandroni	0.9884	0.0116	0.0000
Kosse	Nieuwenhuis	0.0000	1.0000	0.0000
Kosse	Sierra-Loaiza	0.4577	0.4726	0.0697
Kosse	Martinc	0.7971	0.2029	0.0000
Kosse	Veenhoven	0.6519	0.2751	0.0730
Kosse	López-Santillán	0.9523	0.0477	0.0000
Kosse	Hacohen (A)	0.9529	0.0471	0.0000
Kosse	Gopal-Patra	0.9887	0.0112	0.0000
Kosse	Hacohen (B)	0.9523	0.0477	0.0000
Kosse	Stout	0.9532	0.0468	0.0000
Kosse	Von Däniken	0.9886	0.0114	0.0000
Kosse	Schaetti	0.9887	0.0113	0.0000
Kosse	Aragon	0.7583	0.2417	0.0000
Kosse	Bayot	0.9887	0.0113	0.0000
Kosse	Garibo	0.9886	0.0113	0.0000
Kosse	Sezerer	0.9890	0.0110	0.0000
Kosse	Raiyani	0.9886	0.0114	0.0000
Kosse	Sandroni	0.9887	0.0113	0.0000
Nieuwenhuis	Sierra-Loaiza	0.4776	0.3800	0.1423
Nieuwenhuis	Martinc	0.6536	0.3464	0.0000
Nieuwenhuis	Veenhoven	0.6517	0.2759	0.0723
Nieuwenhuis	López-Santillán	0.9537	0.0463	0.0000
Nieuwenhuis	Hacohen (A)	0.9523	0.0477	0.0000
Nieuwenhuis	Gopal-Patra	0.9533	0.0467	0.0000
Nieuwenhuis	Hacohen (B)	0.9887	0.0113	0.0000
Nieuwenhuis	Stout	0.9882	0.0118	0.0000
Nieuwenhuis	Von Däniken	0.9886	0.0114	0.0000
Nieuwenhuis	Schaetti	0.9886	0.0114	0.0000
Nieuwenhuis	Aragon	0.8423	0.1577	0.0000
Nieuwenhuis	Bayot	0.9884	0.0116	0.0000
Nieuwenhuis	Garibo	0.9884	0.0116	0.0000
Nieuwenhuis	Sezerer	0.9886	0.0114	0.0000
Nieuwenhuis	Raiyani	0.9884	0.0116	0.0000
Nieuwenhuis	Sandroni	0.9890	0.0110	0.0000

Team(A)	Team (B)	P(A>B)	P(A=B)	P(A<B)
Sierra-Loaiza	Martinc	0.4786	0.3866	0.1348
Sierra-Loaiza	Veenhoven	0.5299	0.1252	0.3448
Sierra-Loaiza	López-Santillán	0.6292	0.2073	0.1635
Sierra-Loaiza	Hacohen (A)	0.6532	0.2759	0.0709
Sierra-Loaiza	Gopal-Patra	0.6291	0.2084	0.1624
Sierra-Loaiza	Hacohen (B)	0.6514	0.2752	0.0734
Sierra-Loaiza	Stout	0.7589	0.2411	0.0000
Sierra-Loaiza	Von Däniken	0.8445	0.1554	0.0000
Sierra-Loaiza	Schaetti	0.9533	0.0467	0.0000
Sierra-Loaiza	Aragon	0.4803	0.3790	0.1407
Sierra-Loaiza	Bayot	0.9885	0.0114	0.0000
Sierra-Loaiza	Garibo	0.9887	0.0113	0.0000
Sierra-Loaiza	Sezerer	0.9882	0.0118	0.0000
Sierra-Loaiza	Raiyani	0.9884	0.0115	0.0000
Sierra-Loaiza	Sandroni	0.9876	0.0124	0.0000
Martinc	Veenhoven	0.3399	0.3194	0.3407
Martinc	López-Santillán	0.0415	0.9585	0.0000
Martinc	Hacohen (A)	0.4000	0.6000	0.0000
Martinc	Gopal-Patra	0.4734	0.5266	0.0000
Martinc	Hacohen (B)	0.7589	0.2411	0.0000
Martinc	Stout	0.5878	0.4122	0.0000
Martinc	Von Däniken	0.9524	0.0476	0.0000
Martinc	Schaetti	0.9885	0.0115	0.0000
Martinc	Aragon	0.4069	0.5930	0.0000
Martinc	Bayot	0.9879	0.01209	0.0000
Martinc	Garibo	0.9885	0.0115	0.0000
Martinc	Sezerer	0.9885	0.0115	0.0000
Martinc	Raiyani	0.9881	0.0119	0.0000
Martinc	Sandroni	0.9883	0.0117	0.0000
Veenhoven	López-Santillán	0.2767	0.5970	0.1263
Veenhoven	Hacohen (A)	0.4047	0.5953	0.0000
Veenhoven	Gopal-Patra	0.4364	0.5100	0.0536
Veenhoven	Hacohen (B)	0.4061	0.5938	0.0000
Veenhoven	Stout	0.4577	0.4730	0.0693
Veenhoven	Von Däniken	0.8883	0.1117	0.0000
Veenhoven	Schaetti	0.9532	0.0468	0.0000
Veenhoven	Aragon	0.8444	0.1556	0.0000
Veenhoven	Bayot	0.9884	0.0116	0.0000
Veenhoven	Garibo	0.9883	0.0117	0.0000
Veenhoven	Sezerer	0.9883	0.0116	0.0000
Veenhoven	Raiyani	0.9887	0.0113	0.0000
Veenhoven	Sandroni	0.9886	0.0114	0.0000
López-Santillán	Hacohen (A)	0.1537	0.7987	0.0476
López-Santillán	Gopal-Patra	0.0407	0.9593	0.0000
López-Santillán	Hacohen (B)	0.1432	0.8567	0.0000
López-Santillán	Stout	0.5850	0.4149	0.0000
López-Santillán	Von Däniken	0.9525	0.0475	0.0000
López-Santillán	Schaetti	0.9531	0.0468	0.0000
López-Santillán	Aragon	0.4821	0.3152	0.2027
López-Santillán	Bayot	0.9536	0.0464	0.0000
López-Santillán	Garibo	0.9885	0.01151	0.0000
López-Santillán	Sezerer	0.9883	0.01167	0.0000
López-Santillán	Raiyani	0.9885	0.0115	0.0000
López-Santillán	Sandroni	0.9889	0.01109	0.0000

Team(A)	Team (B)	P(A>B)	P(A=B)	P(A<B)
Hacohen (A)	Gopal-Patra	0.1008	0.8539	0.0454
Hacohen (A)	Hacohen (B)	0.0000	1.0000	0.0000
Hacohen (A)	Stout	0.1423	0.8577	0.0000
Hacohen (A)	Von Däniken	0.9533	0.0467	0.0000
Hacohen (A)	Schaetti	0.9542	0.0458	0.0000
Hacohen (A)	Aragon	0.4598	0.4728	0.0674
Hacohen (A)	Bayot	0.9885	0.0115	0.0000
Hacohen (A)	Garibo	0.9883	0.0117	0.0000
Hacohen (A)	Sezerer	0.9886	0.0114	0.0000
Hacohen (A)	Raiyani	0.9884	0.0116	0.0000
Hacohen (A)	Sandroni	0.9886	0.0114	0.0000
Gopal-Patra	Hacohen (B)	0.3238	0.6259	0.0503
Gopal-Patra	Stout	0.2478	0.7031	0.0491
Gopal-Patra	Von Däniken	0.8447	0.1553	0.0000
Gopal-Patra	Schaetti	0.8443	0.1557	0.0000
Gopal-Patra	Aragon	0.4836	0.2177	0.2986
Gopal-Patra	Bayot	0.8450	0.1549	0.0000
Gopal-Patra	Garibo	0.9883	0.0117	0.0000
Gopal-Patra	Sezerer	0.9882	0.0118	0.0000
Gopal-Patra	Raiyani	0.9883	0.0117	0.0000
Gopal-Patra	Sandroni	0.9880	0.0120	0.0000
Hacohen (B)	Stout	0.04087	0.9591	0.0000
Hacohen (B)	Von Däniken	0.8880	0.1120	0.0000
Hacohen (B)	Schaetti	0.8864	0.1136	0.0000
Hacohen (B)	Aragon	0.4749	0.1623	0.3628
Hacohen (B)	Bayot	0.9531	0.0469	0.0000
Hacohen (B)	Garibo	0.9879	0.0121	0.0000
Hacohen (B)	Sezerer	0.9884	0.0116	0.0000
Hacohen (B)	Raiyani	0.9887	0.0113	0.0000
Hacohen (B)	Sandroni	0.9879	0.0121	0.0000
Stout	Von Däniken	0.5863	0.4137	0.0000
Stout	Schaetti	0.7607	0.2393	0.0000
Stout	Aragon	0.4623	0.0777	0.4600
Stout	Bayot	0.9524	0.0476	0.0000
Stout	Garibo	0.9885	0.0115	0.0000
Stout	Sezerer	0.9883	0.0117	0.0000
Stout	Raiyani	0.9881	0.0119	0.0000
Stout	Sandroni	0.9882	0.0118	0.0000
Von Däniken	Schaetti	0.0412	0.9588	0.0000
Von Däniken	Aragon	0.4353	0.0213	0.5434
Von Däniken	Bayot	0.8436	0.1564	0.0000
Von Däniken	Garibo	0.9882	0.01183	0.0000
Von Däniken	Sezerer	0.9881	0.0119	0.0000
Von Däniken	Raiyani	0.9529	0.0471	0.0000
Von Däniken	Sandroni	0.9884	0.0116	0.0000
Schaetti	Aragon	0.4306	0.0213	0.5482
Schaetti	Bayot	0.8420	0.1580	0.0000
Schaetti	Garibo	0.9525	0.0475	0.0000
Schaetti	Sezerer	0.9879	0.0121	0.0000
Schaetti	Raiyani	0.9538	0.0462	0.0000
Schaetti	Sandroni	0.9884	0.0116	0.0000

Team(A)	Team (B)	P(A>B)	P(A=B)	P(A<B)
Aragon	Bayot	0.8447	0.1553	0.0000
Aragon	Garibo	0.8441	0.1559	0.0000
Aragon	Sezerer	0.8698	0.0206	0.1096
Aragon	Raiyani	0.7952	0.0485	0.1563
Aragon	Sandroni	0.8710	0.0202	0.1088
Bayot	Garibo	0.3411	0.3178	0.3411
Bayot	Sezerer	0.5580	0.3878	0.0542
Bayot	Raiyani	0.6305	0.2070	0.1625
Bayot	Sandroni	0.4555	0.4768	0.0677
Garibo	Sezerer	0.4752	0.1601	0.3647
Garibo	Raiyani	0.4810	0.2171	0.3019
Garibo	Sandroni	0.8787	0.0638	0.0575
Sezerer	Raiyani	0.6285	0.2076	0.1639
Sezerer	Sandroni	0.4601	0.4721	0.0678
Raiyani	Sandroni	0.7955	0.0492	0.1553

Appendix C Team Names and Working Notes Authors

In Table A10 the correspondence between team names in TIRA and working notes authors is presented.

Team name	Working note author
aragon18	Aragon & Lopez
bayot18	Bayot & Gonçalves
daneshvar18	Daneshvar
gariboiorts18	Garibo
gouravdas18	Gopal-Patra *et al.*
karlgren18	Karlgren *et al.*
laporte18	Ciccone *et al.*
lopezsantillan18	López-Santillán *et al.*
martinc18	Martinc *et al.*
miller18	Hacohen-Kerner *et al.* (A)
miranda18	Tellez *et al.*
pool18	Stout *et al.*
raiyani18	Raiyani *et al.*
sandroni18	Sandroni-Dias and Paraboni
schaetti18	Schaetti
schuur18	Kosse *et al.*
sierraloaiza18	Sierra-Loaiza & González
snijders18	Veenhoven *et al.*
takahashi18	Takahashi *et al.*
tekir18	Sezerer *et al.*
vaneerden18	Nieuwenhuis & Wilkens
vondaniken18	von Däniken *et al.*
yigal18	Hacohen-Kerner *et al.* (B)

Table A10. Correspondence between TIRA team names and working notes authors.