Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter

Francisco Rangel[1,2] Paolo Rosso[2] Martin Potthast[3] Benno Stein[3]

[1]Autoritas Consulting, S.A., Spain
[2]PRHLT Research Center, Universitat Politècnica de València, Spain
[3]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 task at PAN 2017. The objective of this year is to address gender and language variety identification. For this purpose a corpus from Twitter has been provided for four different languages: Arabic, English, Portuguese, and Spanish. Altogether, the approaches of 22 participants are evaluated.

Malvina - linear svm is good

1 Introduction

The rise of social media provides new models of communication and social relationships. These media allow to hide the real profile of the users who interact and generate information. Therefore, the possibility of knowing social media users' traits on the basis of what they share is a field of growing interest named author profiling. To infer the authors' gender, age, native language, dialects, or personality opens a world of possibilities from the point of view of marketing, forensics, or security. For example from a security viewpoint, to be able to determine the linguistic profile of a person who writes a suspicious or threatening text may provide valuable background information to evaluate the context (and possible reach) of the thread. Moreover, to know the demographics of the author, such as her/his age and gender, or her/his cultural and social context (e.g., native language or/and dialect), with the attempt of profiling potential terrorists [54].

In the Author Profiling task at PAN 2013[1] [48], the identification of age and gender relied on a large corpus collected from social media, both in English and Spanish. In PAN 2014[2] [49], 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 in English only, all documents were provided in both English and Spanish. Note that most of the existing research in computational linguistics [8] and social psychology [43] 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] [50], we included two

[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
[3] http://pan.webis.de/clef15/pan15-web/author-profiling.html

new languages, Italian and Dutch, besides a new subtask on personality recognition. In PAN 2016[4] [52], we investigated the effect of the cross-genre evaluation, that is, when the models are trained on one genre, namely Twitter, and evaluated on another genre different than Twitter.

In PAN 2017[5] we introduce two novelties: (1) the language variety identification together with the gender dimension; and (2) the Arabic and Portuguese languages (besides English and Spanish).

The remainder of this paper is organised 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. Section 5 and 6 discuss results and draw conclusions respectively.

2 Related Work

Pennebaker [44] investigated from a psycholinguistic viewpoint how the use of language varies depending on personal traits such as the author's gender. Concretely, they found that, at least for English, women use the first person of singular more than men because they are more self-conscious, whereas men use more determiners because they speak about concrete things. On the basis of their findings, the authors built LIWC (Linguistic Inquiry and Word Count) [43], one of the most used tools in author profiling. Pioneer researchers such as Argamon *et al.* [8], Holmes & Meyerhoff [23], Burger *et al.* [12], Koppel *et al.* [30] and Schler *et al.* [57] focused mainly on formal texts and blogs, reporting accuracies over 75%-80% in most cases. However, nowadays the investigation is especially focused on social media such as Twitter or Facebook. In this regard, 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 [32, 33, 7]. Recently, the EmoGraph graph-based approach [47] tried to capture how users convey verbal emotions in the morphosyntactic structure of the discourse, obtaining competitive results with the best performing systems at PAN 2013 and demonstrating its robustness against genres and languages at PAN 2014 [46]. Moreover, the authors in [60] investigated on the PAN-AP-2013 dataset a high variety of different features and showed the contribution of information-retrieval-based features in age and gender identification, and in [36] the authors approached the task with 3 million features in a MapReduce configuration, obtaining high accuracies with fractions of processing time. With respect to gender identification in other languages than English and Spanish, it is worth to mention the following investigations. Estival *et al.* [16] focused on Arabic emails and reported accuracies of 72.10%. Alsmearat *et al.* [5] reported an accuracy of 86.4% in Arabic newsletters, and an increase on accuracy up to 94% with an extension of their work [4]. AlSukhni & Alequr [6] reported accuracies of 99.50% by improving a bag-of-words model with the authors' names in Arabic tweets.

This is the first time we have included the language variety identification in the author profiling task. There are a number of investigations with different languages such as

[4] http://pan.webis.de/clef16/pan16-web/author-profiling.html
[5] http://pan.webis.de/clef17/pan17-web/author-profiling.html

English [35], South-Slavic [31], Chinese [24], Persian and Dari [38], or Malay and Indonesian [9]. With respect to Spanish, the authors in [37] investigated the identification among Argentinian, Chilean, Colombian, Mexican, and Spanish in Twitter, reporting accuracies about 60-70% with combinations of n-grams and language models. Also in Spanish, the authors in [51] collected the HispaBlogs[6] corpus with five varieties of Spanish: Argentinian, Chilean, Mexican, Peruvian, and Spanish. They reported an accuracy of 71.1% with a low-dimensionality representation in comparison to 72.2% and 70.8% obtained with Skip-grams and Sentence Vectors [17]. Focusing on Portuguese, the authors in [61] collected 1,000 articles from well-known Brazilian[7] and Portugal[8] newsletters. They combined character and word n-grams and reported accuracies of 99.6% with word unigrams, 91.2% with word bigrams and 99.8% with character 4-grams. With regard to Arabic, Sadat *et al.* [55] reported 98% of accuracy with n-grams in 6 Arabic dialects: Egyptian, Iraqi, Gulf, Maghreb, Levantine, and Sudan. Elfardy & Diab [15] reported 85.5% of accuracy discriminating between Egyptian and Modern Standard Arabic with combinations of content and style-based features. The increasing interest in Arabic dialects identification is attested by the eighteen teams participating in the Arabic subtask of the third DSL track [2][9], and by the 22 participants of this year Author Profiling shared task at PAN.

3 Evaluation Framework

In this section we outline the construction of the corpus while covering particular properties, challenges, and novelties. Finally, the evaluation measures are described.

3.1 Corpus

The focus of this year task is on gender and language variety identification in Twitter. Besides English and Spanish, this year for the first time, we have included Arabic and Portuguese. To create the corpus, we have followed the next 7 steps.

Step 1. Languages and Varieties Selection

The following languages and varieties have been selected:

- Arabic: Egypt, Gulf, Levantine, Maghrebi.[10]
- English: Australia, Canada, Great Britain, Ireland, New Zealand, United States.
- Portuguese: Brazil, Portugal.
- Spanish: Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela.

[6] https://github.com/autoritas/RD-Lab/tree/master/data/HispaBlogs
[7] http://www.folha.uol.com.br
[8] http://www.dn.pt
[9] Its difficulty is backed up by the obtained accuracies of about 50%.
[10] The selection of these varieties responds to previous works [55]. Iraqui was also selected but discarded due to the lack of enough tweets.

Step 2. Tweets per Region Retrieval based on language

For each variety, we have selected the capital (or more populated cities) of the region where this variety is used. The list of cities per variety is shown in Table 1. From the city center, tweets in a radius of 10 kilometers have been retrieved.

Table 1. Cities selected as representative of the language varieties.

Language	Variety	City
Arabic	Egypgt	Cairo
	Gulf	Abu Dhabi, Doha, Kuwait, Manama, Mascate, Riyadh, Sana'a
	Levantine	Amman, Beirut, Damascus, Jerusalem
	Maghrebi	Algiers, Rabat, Tripoli, Tunis
English	Australia	Canberra, Sydney
	Canada	Toronto, Vancouver
	Great Britain	London, Edinburgh, Cardiff
	Ireland	Dublin
	New Zealand	Wellington
	United States	Washington
Portuguese	Brazil	Brasilia
	Portugal	Lisbon
Spanish	Argentina	Buenos Aires
	Chile	Santiago
	Colombia	Bogota
	Mexico	Mexico
	Peru	Lima
	Spain	Madrid
	Venezuela	Caracas

Step 3. Unique Authors Identification

Unique authors have been identified from the previous dataset. For each author, her/his timeline has been retrieved. The timeline provides meta-data such as:

- Full name.
- Location, as a textual description or toponyms.
- Language identified by the author in her/his profile (this language may not correspond to the language used by the author).

Step 4. Authors Selection

For each author, we ensure that the author contains at least 100 tweets with the following conditions:

- Tweets are not retweets.
- Tweets are written in the corresponding language.

Authors who do not accomplish the previous conditions are discarded.

Step 5. Language Variety Annotation

An author is annotated with a corresponding language variety if:

Table 2. Languages and varieties. There are 500 authors per variety and gender, 300 for training and 200 for test. Each author contains 100 tweets.

(AR) Arabic	(EN) English	(ES) Spanish	(PT) Portuguese
Egypt	Australia	Argentina	Brazil
Gulf	Canada	Chile	Portugal
Levantine	Great Britain	Colombia	
Maghrebi	Ireland	Mexico	
	New Zealand	Peru	
	United States	Spain	
		Venezuela	
4,000	6,000	7,000	2,000

- It has been retrieved in the corresponding region.
- At least 80% of the locations provided as meta-data of her/his tweets coincide with some of the toponyms for the corresponding region.

The main assumption is that a person who lives in a region uses this region variety. This implies two assumptions:

- We assume that a person lives in a region when her/his location in all her/his timeline reflects this location. The timeline is up to 3,200 tweets per author, what implies in most cases at least a couple of years, so the assumption is feasible.
- We assume that social media language is dynamic and easily influenced, as opposed to more formal ones such as newsletters. This means that it reflects the everyday language and captures basic social and personality processes of the authors who use it. In this sense, if there is a high number of immigrants or tourists in a region, they may influence the region use of this language, and this may be a valuable clue to detect the possible location of a person.

Step 6. Gender Annotation

Gender annotation has been done in two steps:

- Automatically, with the help of a dictionary of proper nouns (ambiguous nouns have been discarded).
- Manually, by visiting each profile and looking at the photo, description, etc.

Step 7. Corpus construction

The final dataset is balanced in the number of tweets per variety and gender, and in the number of tweets per author:

- 500 tweets per gender and variety.
- 100 tweets per author.

The dataset is divided into training/test in a 60/40 proportion, with 300 authors for training and 200 authors for test. The corresponding languages and varieties are shown in Table 2 along with the total number of authors for each subtask.

3.2 Performance Measures

For evaluation purposes, the accuracy for variety, gender, and joint identification per language is calculated. Then, we average the results obtained per language (Eq. 1).

$$\begin{aligned} \overline{\text{gender}} &= \frac{\text{gender\_AR} + \text{gender\_EN} + \text{gender\_ES} + \text{gender\_PT}}{4} \\ \overline{\text{variety}} &= \frac{\text{variety\_AR} + \text{variety\_EN} + \text{variety\_ES} + \text{variety\_PT}}{4} \\ \overline{\text{joint}} &= \frac{\text{joint\_AR} + \text{joint\_EN} + \text{joint\_ES} + \text{joint\_PT}}{4} \end{aligned} \tag{1}$$

The final ranking is calculated as the average of the above values:

$$\text{ranking} = \frac{\overline{\text{gender}} + \overline{\text{variety}} + \overline{\text{joint}}}{3} \tag{2}$$

3.3 Baselines

To understand the complexity of the subtasks per language and with the aim to compare the performances of the participants approaches, we propose the following baselines:

- *BASELINE-stat.* A statistical baseline that emulates random choice. The baseline depends on the number of classes: two in case of gender identification, and from two to seven in case of variety identification.
- *BASELINE-bow.* This method represents documents as a bag-of-words with the 1,000 most common words in the training set, weighted by absolute frequency of occurrence. The texts are preprocessed as follows: lowercase words, removal of punctuation signs and numbers, and removal of stop words for the corresponding language.
- *BASELINE-LDR* [51]. This method represents documents on the basis of the probability distribution of occurrence of their words in the different classes. The key concept of LDR is a weight, representing the probability of a term to belong to one of the different categories: for gender (female vs. male) and for variety depending on the language (e.g., Brazil vs. Portugal). The distribution of weights for a given document should be closer to the weights of its corresponding category. LDR takes advantage of the whole vocabulary.

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 (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, 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

22 teams participated in the Author Profiling shared task and 20 of them submitted the notebook paper.[11] We analyse their approaches from three perspectives: preprocessing, features to represent the authors' texts and classification approaches.

4.1 Preprocessing

Various participants cleaned the contents to obtain plain text [26, 40, 53]. Most of them removed or normalised Twitter specific elements such as URLs, user mentions or hashtags [18, 1, 27, 29, 39, 41, 53, 56]. Some participants also lowercased the words [18, 27, 29, 41], although in case of [41] the authors did not lowercase the characters. The authors in [1] expanded contractions and the authors in [27, 40] removed stop words. The authors in [53, 40] removed punctuation signs as well as the authors in [56], who also removed numbers and out-of-alphabet words per language. Finally, the authors in [27] removed short tweets.

4.2 Features

Traditionally, in previous editions of the author profiling task at PAN as well as in the referred literature, features used for representing documents have been classified between content and style-based. However, this year for the first time, more participants have employed deep learning techniques. In this sense, it is interesting to differentiate among traditional features and these new methods in order to test their performance in the author profiling task. In this regard, the authors in [25, 29, 58, 45] represented documents with word embeddings, whereas in [18] character embeddings have been used. The authors in [41] mixed both word and character embeddings and the authors in [45] also used traditional features such as tf-idf n-grams.

As traditional features, character and word n-grams have been widely used by the participants [40, 3, 27, 39, 53, 42, 56, 14]. For example, the authors in [3] used character n-grams with n between 2 and 7, and the authors in [27] used word n-grams with values between 1 and 3 for the n. The authors in [39] combined character and word n-grams with values of 3-4 for typed characters, 3-7 for untyped characters and 2-3 for words, respectively. Similarly, the authors in [14] combined character and word n-grams with values of n of 1-6 and 1-2 respectively. In case of the authors of [19], words with frequency between 2 and 25 have been used, whereas in [11] the authors combined most discriminative words per class with slang words, locations, brands and stylistic patterns.

[11] Ganesh [19] and Bouazizi [11] teams did not submit a notebook paper, but sent us a brief description of their approaches.

In [45, 56] tf-idf n-grams have been combined respectively with word embeddings, and with beginning and ending character 2-grams. Finally, the authors in [42] used high order character n-grams.

With respect to content features, the most commonly used have been top n terms by gain ratio [28], bag-of-words [1, 59], the 100 most discriminant words per class from a list of 500 topic words [26], LSA [27], specific lists of words for language variety [40]. Style features have been also used by some participants. For example, ratios of links, hashtags or user mentions [3, 53], character flooding [3, 40, 53], and emoticons or/and laugher expressions [1, 40]. Finally, the authors in [39] combined domain names with different kinds of n-grams.

Emotional features have been used by the authors in [1], who combined emotions, appraisal, admiration, positive/negative emoticons, and positive/negative words, and by the authors in [40] who used emojis and sentiment words. The authors in [34] used a variation of their second-order representation [33] based on user-document relationships. Finally in [10], the authors who obtained the best overall result, used a combination of character n-grams (with n between 3 and 5) and tf-idf word n-grams (with n between 1 and 2).

4.3 Classification Approaches

Most of the participants approached the task with traditional machine learning algorithms such as logistic regression [25, 40, 45, 42], SVMs [3, 27, 34, 39, 59, 10, 53, 14, 19] and Random Forest [11]. Meanwhile most participants used these algorithms alone, authors in [45, 14] ensembled different configurations. The authors in [27] used SVMs for variety identification and Naive Bayes for gender identification. Three teams used distance-based methods [1, 26, 28].

With respect to deep learning methods, the authors in [29] applied Recurrent Neural Networks (RNN), whereas the authors in [56, 58] used Convolutional Neural Networks (CNN). The authors in [41] explored both approaches (RNN and CNN) besides attention mechanism, max-pooling layer, and fully-connected layer. The authors in [45] combined traditional logistic regression with a Gaussian process trained with word embeddings. Finally, the authors in [18] applied Deep Averaging Networks.

5 Evaluation and Discussion of the Submitted Approaches

We divided the evaluation in two steps, providing an early bird option for those participants who wanted to receive some feedback. There were 22 submissions for the final evaluation. We show results separately for the evaluation in each subtask (gender, language variety, and joint identification), as well as we analyse the special case of the English language in a coarse-grained grouping.

5.1 Gender Evaluation

In this section we analyse the results for the gender identification subtask. As can be seen in Table 3, the best results have been obtained for the Portuguese language with a

maximum accuracy of 87% [41] and an average result of 78%, about 7 and 3 percentage points over the rest. However, results for the four languages are very similar in terms of their distribution, as can be seen in Figure 1. The average values range between 72.10% in case of Arabic and 78% in case of Portuguese. The most similar results among authors have been obtained in English, although there are two outliers with lower results than the rest: the authors in [1] with an accuracy of 54.13%, obtained with combinations of content and style-based features learned with a distance-based method, and Bouazizi's team that achieved 61.21% with Random Forest and combinations of discriminative words, slang, locations, brands and stylistic patterns. In case of Portuguese there is also an outlier with lower accuracy than the rest, corresponding to the authors in [26] with 61% of accuracy, obtained with top 100 most discriminant words per class and a distance-based algorithm.

The best results per language have been obtained by the following teams: In Arabic, the authors in [40] approached the task with combinations of character, word and POS n-grams with emojis, character flooding, sentiment words, and specific lists of words per variety, training their models with logistic regression and obtaining an accuracy of 80.31%. In case of English, the best result of 82.33% has been obtained by the authors in [10]. The authors approached the task with combinations of character and tf-idf word n-grams trained with an SVM. With respect to Portuguese, the best result has been obtained by the authors in [41]. They achieved 87% accuracy with a deep learning approach combining word and character embeddings with CNN, RNN, attention mechanism, max-pooling layer, and fully-connected layer. In case of Spanish, the authors who obtained the best result of 83.21% are the same that obtained the best result in English.

With respect to the provided baselines, especially in case of BOW and LDR that both utilize information from the contents of the documents, we can observe that their results are below the mean. As seen in previous editions of the task and according to previous investigations [44, 47], gender discrimination is more related to how things are said than to what it is said. In this sense, the best resulting approaches took advantadge from both the style and contents with combinations of n-grams and other content and style-based features, as well as in case of Portuguese with deep representations.

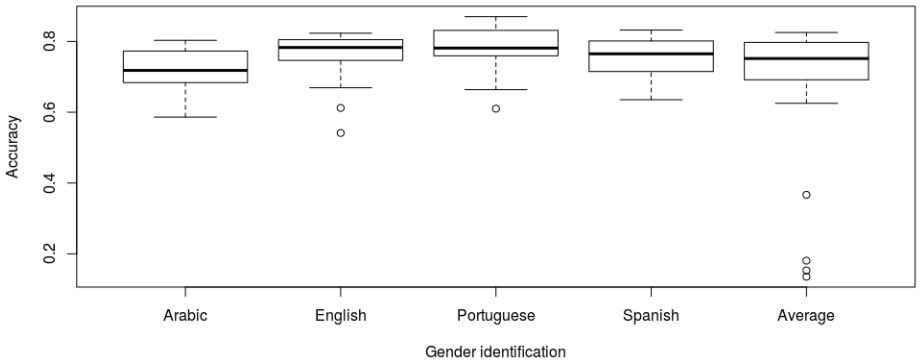

Figure 1. Distribution of results for gender identification in the different languages.

Table 3. Gender results.

Ranking	Team	Arabic	English	Portuguese	Spanish	Average
1	Basile et al.	0.8006	0.8233	0.8450	0.8321	0.8253
2	Martinc et al.	0.8031	0.8071	0.8600	0.8193	0.8224
3	Miura et al.	0.7644	0.8046	0.8700	0.8118	0.8127
4	Tellez et al.	0.7838	0.8054	0.8538	0.7957	0.8097
5	Lopez-Monroy et al.	0.7763	0.8171	0.8238	0.8014	0.8047
6	Poulston et al.	0.7738	0.7829	0.8388	0.7939	0.7974
7	Markov et al.	0.7719	0.8133	0.7863	0.8114	0.7957
8	Ogaltsov & Romanov	0.7213	0.7875	0.7988	0.7600	0.7669
9	Franco-Salvador et al.	0.7300	0.7958	0.7688	0.7721	0.7667
10	Sierra et al.	0.6819	0.7821	0.8225	0.7700	0.7641
11	Kodiyan et al.	0.7150	0.7888	0.7813	0.7271	0.7531
12	Ciobanu et al.	0.7131	0.7642	0.7713	0.7529	0.7504
13	Ganesh	0.6794	0.7829	0.7538	0.7207	0.7342
	LDR-baseline	0.7044	0.7220	0.7863	0.7171	0.7325
14	Schaetti	0.6769	0.7483	0.7425	0.7150	0.7207
15	Kocher & Savoy	0.6913	0.7163	0.7788	0.6846	0.7178
16	Kheng et al.	0.6856	0.7546	0.6638	0.6968	0.7002
17	Ignatov et al.	0.6425	0.7446	0.6850	0.6946	0.6917
	BOW-baseline	0.5300	0.7075	0.7812	0.6864	0.6763
18	Khan	0.5863	0.6692	0.6100	0.6354	0.6252
	STAT-baseline	0.5000	0.5000	0.5000	0.5000	0.5000
19	Ribeiro-Oliveira et al.	0.7013	-	0.7650	-	0.3666
20	Alrifai et al.	0.7225	-	-	-	0.1806
21	Bouazizi	-	0.6121	-	-	0.1530
22	Adame et al.	-	0.5413	-	-	0.1353
	Min	0.5863	0.5413	0.6100	0.6354	0.1353
	Q1	0.6847	0.7474	0.7594	0.7164	0.6938
	Median	0.7181	0.7829	0.7813	0.7650	0.7518
	Mean	0.7210	0.7571	0.7800	0.7553	0.6588
	SDev	0.0560	0.0729	0.0690	0.0554	0.2260
	Q3	0.7724	0.8048	0.8313	0.8000	0.7970
	Max	0.8031	0.8233	0.8700	0.8321	0.8253

5.2 Language Variety Evaluation

In this section we analyse the results for the language variety identification subtask. As can be seen in Table 4, the best results have been obtained for the Portuguese language with a maximum accuracy of 98.50% [59] and an average result of 97.31%.

The best results per language have been obtained by the following teams: In Arabic and Spanish, the authors in [10] obtained 83.13% and 96.21% of accuracy respectively, approaching the task with combinations of character and tf-idf word n-grams and an SVM. With respect to English and Portuguese, the authors in [59] obtained 90.04% and 98.50% with an SVM.

The provided LDR baseline obtained almost the best result in all languages and obtained the best overall result. Taking into account that basically this representation measures the use of words per class, we can conclude that the identification of language varieties is highly dependent to the usage of words. This is also supported by the approaches used by the best performing teams.

Table 4. Language variety results.

Ranking	Team	Arabic	English	Portuguese	Spanish	Average
	LDR-baseline	0.8250	0.8996	0.9875	0.9625	0.9187
1	Basile et al.	0.8313	0.8988	0.9813	0.9621	0.9184
2	Tellez et al.	0.8275	0.9004	0.9850	0.9554	0.9171
3	Martinc et al.	0.8288	0.8688	0.9838	0.9525	0.9085
4	Markov et al.	0.8169	0.8767	0.9850	0.9439	0.9056
5	Lopez-Monroy et al.	0.8119	0.8567	0.9825	0.9432	0.8986
6	Miura et al.	0.8125	0.8717	0.9813	0.9271	0.8982
7	Sierra et al.	0.7950	0.8392	0.9850	0.9450	0.8911
8	Schaetti	0.8131	0.8150	0.9838	0.9336	0.8864
9	Poulston et al.	0.7975	0.8038	0.9763	0.9368	0.8786
10	Ogaltsov & Romanov	0.7556	0.8092	0.9725	0.8989	0.8591
11	Ciobanu et al.	0.7569	0.7746	0.9788	0.8993	0.8524
12	Kodiyan et al.	0.7688	0.7908	0.9350	0.9143	0.8522
13	Kheng et al.	0.7544	0.7588	0.9750	0.9168	0.8513
14	Franco-Salvador et al.	0.7656	0.7588	0.9788	0.9000	0.8508
15	Kocher & Savoy	0.7188	0.6521	0.9725	0.7211	0.7661
16	Ganesh	0.7144	0.6021	0.9650	0.7689	0.7626
17	Ignatov et al.	0.4488	0.5813	0.9763	0.8032	0.7024
	BOW-baseline	0.3394	0.6592	0.9712	0.7929	0.6907
18	Khan	0.5844	0.2779	0.9063	0.3496	0.5296
19	Ribeiro-Oliveira et al.	0.6713	-	0.9850	-	0.4141
	STAT-baseline	0.2500	0.1667	0.5000	0.1429	0.2649
20	Alrifai et al.	0.7550	-	-	-	0.1888
21	Bouazizi	-	0.3725	-	-	0.0931
22	Adame et al.	-	0.1904	-	-	0.0476
	Min	0.4488	0.1904	0.9063	0.3496	0.0476
	Q1	0.7455	0.6396	0.9738	0.8990	0.7175
	Median	0.7672	0.7973	0.9788	0.9220	0.8523
	Mean	0.7514	0.7150	0.9731	0.8707	0.7215
	SDev	0.0936	0.2098	0.0198	0.1466	0.2798
	Q3	0.8126	0.8597	0.9838	0.9437	0.8964
	Max	0.8313	0.9004	0.9850	0.9621	0.9184

As can be seen in Figure 2, the distribution of the results is varies significantly with the language. In case of Portuguese, most approaches obtained very similar results, with a difference of only 1% in the interquartile range. The distribution of results in case of Arabic and Spanish are similar, with an interquartile range of 6.71% and 4.47% respectively, although results for Spanish are higher (an average of 87.07% over 75.14%). The most sparse distribution occurs for English, where the interquartile range embraces 22.01% with the lowest average accuracy of 71.50%. However, this lowest average can be caused by the outlier [1], who obtained an accuracy of 19.04% with a distance-based approach with combinations of bag-of-words and emotional features.

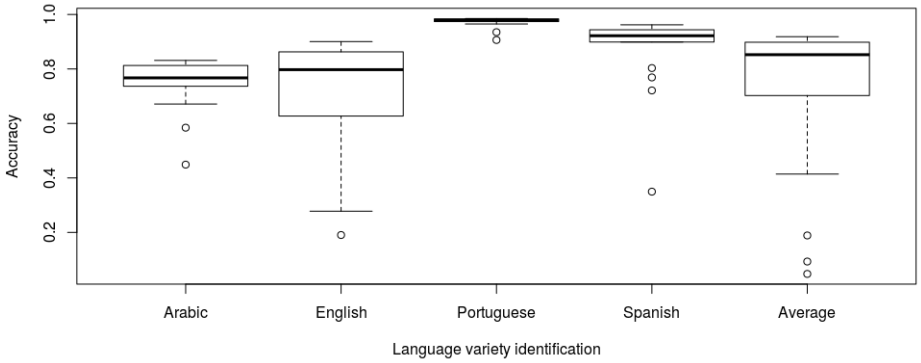

Figure 2. Distribution of results for language variety identification in the different languages.

5.3 Confusion Among Language Varieties

In this section the confusion among varieties of the same language is analysed. All participants' results have been analysed together; the Figures 3, 4, 5 and 6 show the confusion matrix for Arabic, English, Portuguese, and Spanish respectively.

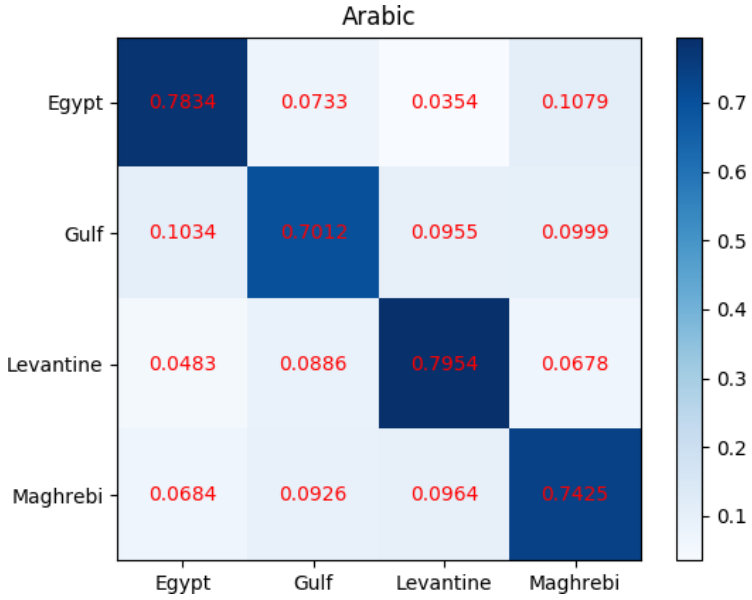

Figure 3. Confusion matrix for Arabic varieties.

In Figure 3 the confusion matrix for Arabic languages is shown: the overall confusion among languages is at most 10.79% in case of Egypt to Maghrebi, and 10.34% in case of Gulf to Egypt. The rest of the errors are below 10%, with the lowest value for Egypt to Levantine (3.54%), or Levantine to Egypt (4.83%). The highest accuracy was obtained for the Levantine variety (79.54%) that, together with the previous insights, show us that this variety seems to be the less difficult to be identified. On the contrary,

the Gulf variety is the most difficult one, with an overall accuracy of 70.12% and the highest confusions to other varieties.

Figure 4 shows the confusion matrix for the English language. The highest confusion is from United States to Canada (11.80%), Ireland to Great Britain (11.81%), and Australia to New Zealand (10.61%). These errors correspond to varieties with a close geographical situation. The New Zealand variety is the less difficult to be identified (82.04%), whereas the most difficult is Australia (65.48%). It is noteworthy that both varieties are geographically close, but the error from New Zealand to Australia is very low (3.46%). The lowest confusion is from New Zealand to Ireland (1.7%), United States to Ireland (2.24%), Ireland to Australia (2.50%), or Canada to Ireland (2.59%). As can be concluded, the closer geographically two varieties are, the higher the confusion between them is.

Figure 4. Confusion matrix for English varieties.

The confusion between the two varieties of Portuguese is shown in Figure 5. The Brazilian variety is less difficult to be identified than the Portuguese one. The error from Portugal to Brazil is 4.45%, whereas the error in the other direction is lower than 1%.

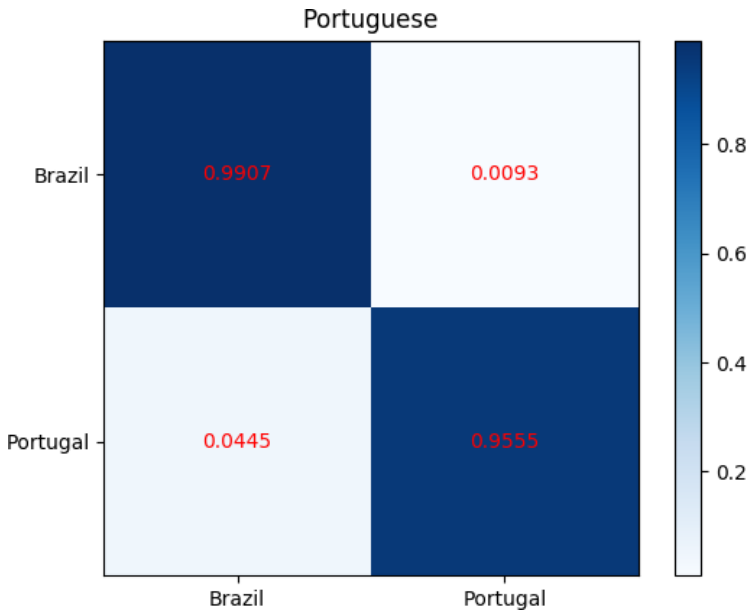

Figure 5. Confusion matrix for Portuguese varieties.

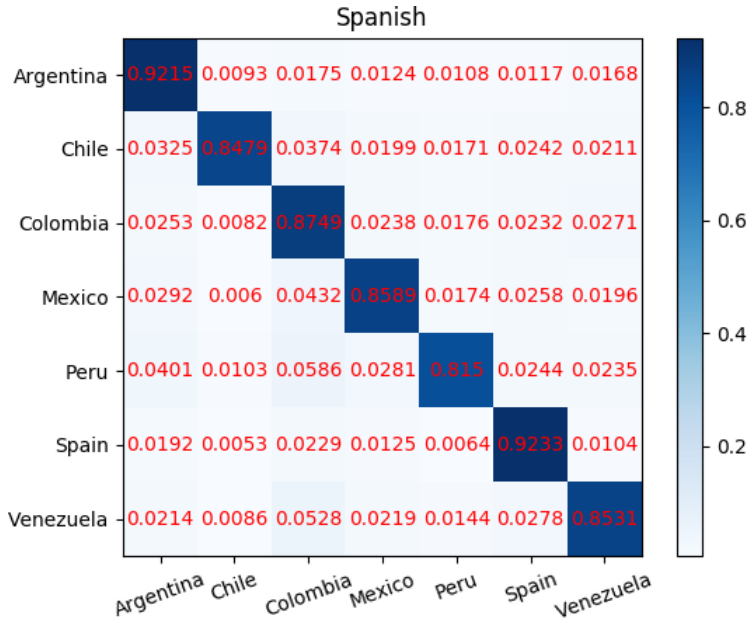

Figure 6. Confusion matrix for Spanish varieties.

In case of Spanish the confusion matrix among varieties is shown in Figure 6. It can be observed that the varieties from Argentina and Spain are the less difficult to be identified, with accuracies of 92.15% and 92.33% respectively. On the contrary, the most difficult variety is Peru, where the accuracy drops to 81.5%, followed by Chile (84.79%).

The highest confusion occurs from Peru, Venezuela, and Mexico to Colombia, with errors of 5.86%, 5.28%, and 4.32% respectively, followed by Peru to Argentina (4.01%), Chile to Colombia (3.74%), and Chile to Argentina (3.25%). Colombia is the variety destination of the majority of the confusions, and, again the geographical distribution of varieties has some relationship with the confusion between them.

5.4 Coarse-Grained Evaluation of English Varieties

Due to the confusion among English varieties that share geographical proximitiy, we have combined some of them in order to analyse how the participants' approaches performed. In particluar, we formed the following groups:

- American English: United States and Canada.
- European English: Great Britain and Ireland.
- Oceanic English: Australia and New Zealand.

Table 5. Coarse-grained English variety identification results.

Ranking	Team	Coarse-Grained	Fine-Grained	Difference
1	Basile et al.	0.9429	0.8988	0.0441
2	Tellez et al.	0.9379	**0.9004**	**0.0375**
3	Markov et al.	0.9292	0.8767	0.0525
4	Miura et al.	0.9279	0.8717	0.0562
5	Martinc et al.	0.9238	0.8688	0.0550
6	Lopez-Monroy et al.	0.9167	0.8567	0.0600
7	Sierra et al.	0.9004	0.8392	0.0612
8	Schaetti	0.8863	0.8150	0.0713
9	Ogaltsov & Romanov	0.8754	0.8092	0.0662
10	Poulston et al.	0.8746	0.8038	0.0708
11	Kodiyan et al.	0.8663	0.7908	0.0755
12	Franco-Salvador et al.	0.8654	0.7588	0.1066
13	Ciobanu et al.	0.8504	0.7746	0.0758
14	Kheng et al.	0.8388	0.7588	0.0800
15	Kocher & Savoy	0.7696	0.6521	0.1175
16	Ignatov et al.	0.7296	0.5813	0.1483
17	Ganesh	0.7238	0.6021	0.1217
18	Bouazizi	0.5217	0.3725	0.1492
19	Khan	0.4533	0.2779	**0.1754**
20	Adame et al.	0.3583	0.1904	0.1679

In Table 5 the results for this grouping (coarse-grained) in comparison to the original varieties (fine-grained) are shown. As can be seen, results are better in case of coarse-grained evaluation, although the difference is higher with systems that performed worst in the fine-grained evaluation. For example, in case of Khan [26], the increase in accuracy is 17.54%, whereas in case of Tellez *et al.* [59] the difference is only 3.75%. Although the statistical test indicates that the differences are significant, such differences are not very high in case of the best performing teams with values around 5% of improvement.

5.5 The Impact of the Gender on the Language Variety Identification

In this section we analyse the impact of the gender on language variety identification, depending on the native language. To carry out this analysis, we have aggregated the results obtained by all participants (without the baselines). Table 6 shows the accuracies obtained in language variety identification depending on the gender. As can be seen, except in the case of Spanish, it is easier to identifiy the variety properly in case of females, although only in case of Arabic and Portuguese this difference is statistically significant. In case of Spanish and English, the difference in accuracy between genders is about 5% and 2% respectively, whereas in Arabic and Portuguese is about 7% and 2%.

Table 6. Language variety identification accuracies per gender (* indicates a significant difference according to the Student t-test).

Language	Female	Male	Difference
Arabic*	0.7909	0.7203	0.0706
English	0.7190	0.7168	0.0022
Portuguese*	0.9829	0.9633	0.0196
Spanish	0.8680	0.8733	-0.0053

5.6 The Difficulty of Gender Identification Depending on the Language Variety

In this section, the difficulty of identifying the author's gender depending on the language variety is analysed. We have followed the same methodology by accumulating the results of all participants without the baselines. The results are shown in Table 7, where statistical significance is marked with an asterisk.

In case of Arabic, females are easier to be identified in all varieties except in case of Maghrebi, where in addition the difference between genders is the highest, with more than 15% of accuracy. Only in case of Levantine the difference is not statistically significant, with a difference of about 2%. With respect to English, there is not an easier gender but it depends on the variety. However, only in case of Australia, New Zealand and United States these diferences are significant, with values of about 5%, 6% and a noteworthy 14%, respectively. Something similar happens with Spanish, where the differences depend on the variety without a predominantly easier gender. In this language, there are four out seven varieties with significant differences. The highest differences occur with Argentina (14%) and Venezuela (10%), whereas with Chile and Spain these differences are minor (about 6% and 7% respectively). In case of Portuguese there is a clear and significant difference among genders in favour of females, where they are easier to be distinguished independently of the variety and with an increase in accuracy of more than 12% and 14%.

Table 7. Gender identification accuracies per language variety (* indicates a significant difference according to a Student t-test).

Language	Variety	Female	Male	Difference	Average
Arabic	Egypt*	0.7426	0.6847	0.7137	0.0579
	Gulf*	0.7932	0.6937	0.7435	0.0995
	Levantine	0.7324	0.7108	0.7216	0.0216
	Maghrebi*	0.6345	0.7847	0.7096	-0.1502
English	Australia*	0.6871	0.7367	0.7119	-0.0496
	Canada	0.7452	0.7412	0.7432	0.0040
	Great Britain	0.7859	0.8063	0.7961	-0.0204
	Ireland	0.8025	0.8015	0.8020	0.0010
	New Zealand*	0.7514	0.8102	0.7808	-0.0588
	United States*	0.6581	0.7962	0.7272	-0.1381
Portuguese	Brazil*	0.8037	0.6834	0.7436	0.1203
	Portugal*	0.8895	0.7432	0.8164	0.1463
Spanish	Argentina*	0.8525	0.7097	0.7811	0.1428
	Chile*	0.7528	0.8153	0.7841	-0.0625
	Colombia	0.7836	0.7708	0.7772	0.0128
	Mexico	0.7028	0.7242	0.7135	-0.0214
	Peru	0.7900	0.7794	0.7847	0.0106
	Spain*	0.7003	0.7697	0.7350	-0.0694
	Venezuela*	0.6603	0.7625	0.7114	-0.1022

5.7 Gender and Language Variety Joint Evaluation

In this section a summary of the joint results per language is shown. Table 9 shows the users' performances when both gender and language variety are properly detected per language. We can observe that the best results were achieved in Portuguese (85.75%), followed by Spanish (80.36%), English (74.29%) and Arabic (68.31%). The difference on accuracy among languages is very significant. Most of the participants obtained better results than two of the baselines (BOW and statistical), although only 8 participants outperformed the LDR baseline. Furthermore, in case of Portuguese only 9 teams outperformed the bag-of-words baseline, showing the power of simple words to discriminate among varieties and genders in that language. On the contrary, this baseline shows its inefficiency in case of Arabic, where the accuracy drops to values close to the statistical baseline.

Looking at Figure 7, it can be observed that there are a couple of outliers obtaining lower results. For Arabic, two participants obtained significantly worse results than the rest. The authors in [25] approached the task with word embeddings and logistic regression, obtaining 28.75% accuracy. The authors in [26] obtained 36.50% with the 100 most discriminant words per class and a distance-based algorithm. This approach also obtained the outlier results in Portuguese (54.88%) and Spanish (21.89%). With regard to English, the authors in [1] obtained 10.17% with combinations of bag-of-words and series of emotional features learned with a distance-based approach.

Table 8. Joint results.

Ranking	Team	Arabic	English	Portuguese	Spanish	Average
1	Basile et al.	0.6831	0.7429	0.8288	0.8036	0.7646
2	Martinc et al.	0.6825	0.7042	0.8463	0.7850	0.7545
3	Tellez et al.	0.6713	0.7267	0.8425	0.7621	0.7507
4	Miura et al.	0.6419	0.6992	0.8575	0.7518	0.7376
5	Lopez-Monroy et al.	0.6475	0.7029	0.8100	0.7604	0.7302
6	Markov et al.	0.6525	0.7125	0.7750	0.7704	0.7276
7	Poulston et al.	0.6356	0.6254	0.8188	0.7471	0.7067
8	Sierra et al.	0.5694	0.6567	0.8113	0.7279	0.6913
	LDR-baseline	0.5888	0.6357	0.7763	0.6943	0.6738
9	Ogaltsov & Romanov	0.5731	0.6450	0.7775	0.6846	0.6701
10	Franco-Salvador et al.	0.5688	0.6046	0.7525	0.7021	0.6570
11	Kodiyan et al.	0.5688	0.6263	0.7300	0.6646	0.6474
12	Ciobanu et al.	0.5619	0.5904	0.7575	0.6764	0.6466
13	Schaetti	0.5681	0.6150	0.7300	0.6718	0.6462
14	Kheng et al.	0.5475	0.5704	0.6475	0.6400	0.6014
15	Ganesh	0.5075	0.4713	0.7300	0.5614	0.5676
16	Kocher & Savoy	0.5206	0.4650	0.7575	0.4971	0.5601
	BOW-baseline	0.1794	0.4713	0.7588	0.5561	0.4914
17	Ignatov et al.	0.2875	0.4333	0.6675	0.5593	0.4869
18	Khan	0.3650	0.1900	0.5488	0.2189	0.3307
19	Ribeiro-Oliveira et al.	0.4831	-	0.7538	-	0.3092
20	Alrifai et al.	0.5638	-	-	-	0.1410
	STAT-baseline	0.1250	0.0833	0.2500	0.0714	0.1324
21	Bouazizi	-	0.2479	-	-	0.0620
22	Adame et al.	-	0.1017	-	-	0.0254
	Min	0.2875	0.1017	0.5488	0.2189	0.0254
	Q1	0.5408	0.4697	0.7300	0.6462	0.5052
	Median	0.5688	0.6202	0.7575	0.6934	0.6470
	Mean	0.5650	0.5566	0.7601	0.6658	0.5552
	SDev	0.1010	0.1854	0.0768	0.1400	0.2308
	Q3	0.6433	0.7001	0.8151	0.7582	0.7224
	Max	0.6831	0.7429	0.8575	0.8036	0.7646

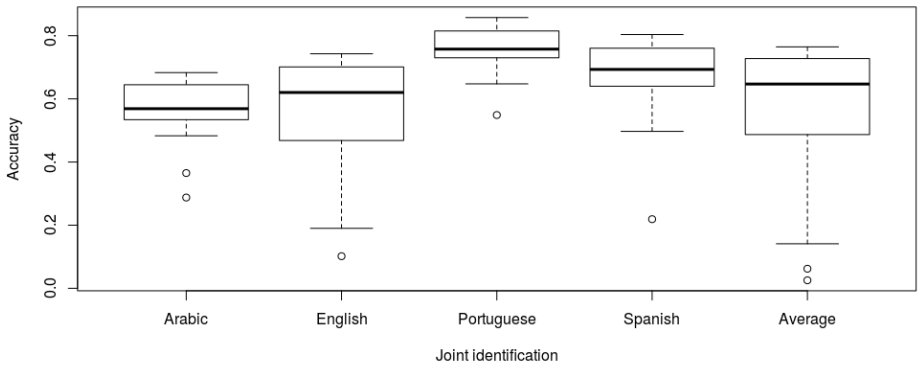

Figure 7. Distribution of results for joint identification in the different languages.

5.8 Final Ranking and Best Results

In Table 9 the final ranking is shown. The values shown in each column correspond to the average of the accuracies for this subtask among the different languages. The ranking shows that the best overall result (85.99%) has been obtained by Basile *et al.* [10], who used an SVM trained with combinations of character and tf-idf n-grams, followed by Martinc *et al.* [40] (85.31%), who used logistic regression with combinations of character, word, and POS n-grams, emojis, sentiments, character flooding, and lists of words per variety. The third best result has been obtained by Tellez *et al.* [59] (85.09%) with an SVM. These three best results are not significantly different.

Table 9. Global ranking as average of each language average for subtask.

Ranking	Team	Gender	Variety	Joint	Average
1	Basile et al.	0.8253	0.9184	0.8361	0.8599
2	Martinc et al.	0.8224	0.9085	0.8285	0.8531
3	Tellez et al.	0.8097	0.9171	0.8258	0.8509
4	Miura et al.	0.8127	0.8982	0.8162	0.8424
5	Lopez-Monroy et al.	0.8047	0.8986	0.8111	0.8381
6	Markov et al.	0.7957	0.9056	0.8097	0.8370
7	Poulston et al.	0.7974	0.8786	0.7942	0.8234
8	Sierra et al.	0.7641	0.8911	0.7822	0.8125
	LDR-baseline	0.7325	0.9187	0.7750	0.8087
9	Ogaltsov & Romanov	0.7669	0.8591	0.7653	0.7971
10	Franco-Salvador et al.	0.7667	0.8508	0.7582	0.7919
11	Schaetti	0.7207	0.8864	0.7511	0.7861
12	Kodiyan et al.	0.7531	0.8522	0.7509	0.7854
13	Ciobanu et al.	0.7504	0.8524	0.7498	0.7842
14	Kheng et al.	0.7002	0.8513	0.7176	0.7564
15	Ganesh	0.7342	0.7626	0.6881	0.7283
16	Kocher & Savoy	0.7178	0.7661	0.6813	0.7217
17	Ignatov et al.	0.6917	0.7024	0.6270	0.6737
	BOW-baseline	0.6763	0.6907	0.6195	0.6622
18	Khan	0.6252	0.5296	0.4952	0.5500
19	Ribeiro-Oliveira et al.	0.3666	0.4141	0.3092	0.3633
	STAT-baseline	0.5000	0.2649	0.2991	0.3547
20	Alrifai et al.	0.1806	0.1888	0.1701	0.1798
21	Bouazizi	0.1530	0.0931	0.1027	0.1163
22	Adame et al.	0.1353	0.0476	0.0695	0.0841
	Min	0.1353	0.0476	0.0695	0.0841
	Q1	0.6938	0.7175	0.6406	0.6857
	Median	0.7518	0.8523	0.7510	0.7857
	Mean	0.6588	0.7215	0.6427	0.6743
	SDev	0.2260	0.2798	0.2472	0.2502
	Q3	0.7970	0.8964	0.8058	0.8336
	Max	0.8253	0.9184	0.8361	0.8599

As can be seen in Figure 8, there are several outliers with lower results than the rest. These outliers correspond in almost all tasks with the same authors, who applied SVMs trained with style-based features such as character flooding or different ratios of

hashtags and mentions [53, 3], a distance-based method with a high variety of emotional features combined with bag-of-words [1].

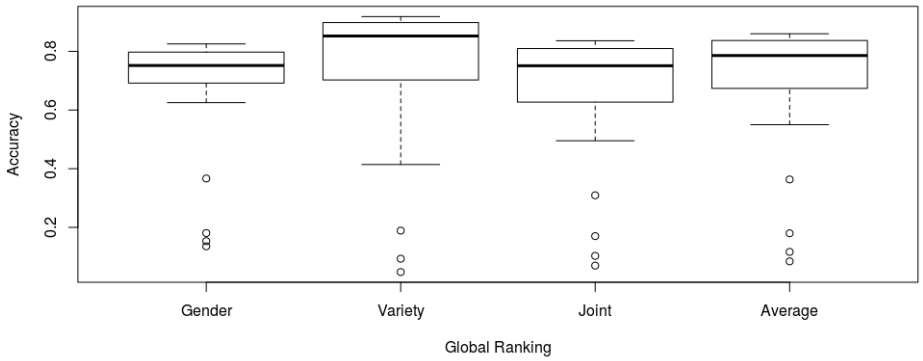

Figure 8. Distribution of results in the different tasks.

In Table 10 the best results per language and task are shown. We can observe that for both the gender and variety subtasks, the best results were achieved in Portuguese, followed by Spanish, English and Arabic. In case of gender identification, the accuracies are between 80.31% in case of Arabic and 87% in case of Portuguese, whereas the difference is higher for language variety identification, where the worst results obtained in Arabic is 83.13% (4 varieties), against a 98.38% obtained in Portuguese (2 varieties). Results for Spanish (7 varieties) (96.21%) are close to Portuguese, while in English (6 varieties) they fall to 89.88%.

Table 10. Best results per language and task.

Language	*Joint*	Gender	Variety
Arabic	0.6831	0.8031	0.8313
English	0.7429	0.8233	0.8988
Spanish	0.8036	0.8321	0.9621
Portuguese	0.8575	0.8700	0.9838

6 Conclusion

In this paper we presented the results of the 5th International Author Profiling Shared Task at PAN 2017 within CLEF 2017. The participants had to identify gender and language variety from Twitter authors collected in four different languages: Arabic, English, Portuguese, and Spanish.

The participants used different feature sets to approach the problem: content-based (among others: bag of words, word n-grams, term vectors, named entities, dictionary words, slang words, contractions, sentiment words) and stylistic-based (among others: frequencies, punctuations, POS, Twitter specific elements, readability measures). For the first time, deep learning approaches have been used: Recurrent Neural Networks,

Convolutional Neural Networks, as well as word and character embeddings. It is difficult to highlight the contribution of each particular feature since the participants used many of them. However, deep learning approaches did not obtain the best results.

In both subtasks as well as in the joint identification, the best results have been obtained for Portuguese. In case of gender identification, average results are quite similar among languages, with the lowest result for Arabic (72.10%), followed by Spanish (75.53%), English (75.71%), and Portuguese (78.00%). In case of language variety the worst average result has been obtained for English (71.50%), followed by Arabic (75.14%), Spanish (87.07%), and Portuguese (97.31%). However, in this case the difference between the worst and the best results is much higher (25.81% vs. 5.9%).

By analysing the error when discriminating among varieties, we found interesting facts: For example, in Arabic, the most difficult to be identified is the Gulf variety, whereas the easiest one is the Levantine. In case of English, the highest confusion occurs with varieties that share the same geographical location: America, Europe, or Oceania. In case of Portuguese, where only two varieties are analysed, the asymmetry in the confusion matrix is noteworthy: most errors occur when Portuguese variety is confused with Brazilian variety (4.45%), unlike when Brazilian is confused with Portugal variety (0.93%). In case of Spanish, most confusions refer to Colombia, although the most confused variety is the from Peru. On the contrary, the easiest varieties to be identified are from Argentina and Spain.

Due to the geographical impact on the English varieties identification, we have carried out a coarse-grained evaluation by combining varieties per continent: America, Europe, and Oceania. Although results increase with statistical significance, the differences are not very high in case of the best performing approaches (3.75%) .

We have analysed the impact of the gender in the language variety identification, showing that in Arabic and Portuguese the difference between genders is statistically significant. Similarly, we have analysed the difficulty of gender identification depending on the language variety, obtaining different insights with respect to both the easiest gender to be identified and the significance of these results. For example, for most Arabic varieties, females are less difficult to be identified (except for Maghrebi), as well as in case of Portuguese. In case of Spanish and English, both genders appear to be not so difficult depending on the variety.

Th joint results show a similar distribution than the final evaluation. In this regard, Basile *et al.* [10], Martinc *et al.* [40], and Tellez *et al.* [59] obtained the best result, without significant difference among them. They approached the task with SVMs trained with combinations of character and tf-idf n-grams, logistic regression with combinations of character, word and POS n-grams, emojis, sentiments, character flooding, and lists of words per variety, learned with an SVM.

Acknowledgements

Our special thanks go to all of PAN's participants, and to MeaningCloud[12] for sponsoring also this edition of the author profiling shared task award. The first author acknowledges the SomEMBED TIN2015-71147-C2-1-P MINECO research project. The work

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

on the data in 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). The statements made herein are solely the responsibility of the first two authors.

Bibliography

[1] Yaritza Adame, Daniel Castro, Reynier Ortega, and Rafael Mu noz. Author profiling, instance-based similarity classification. In Cappellato et al. [13].

[2] Ahmed Ali, Peter Bell, and Steve Renals. Automatic dialect detection in arabic broadcast speech. In *Interspeech*, 2015.

[3] Khaled Alrifai, Ghaida Rebdawi, and Nada Ghneim. Arabic tweeps gender and dialect prediction. In Cappellato et al. [13].

[4] 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.

[5] 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.

[6] 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.

[7] 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. 2015.

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

[9] Ranaivo-Malançon Bali. Automatic identification of close languages–case study: malay and indonesian. *ECTI Transaction on Computer and Information Technology*, 2(2):126–133, 2006.

[10] Agenlo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. Is there life beyond n-grams? a simple svm-based author profiling system. In Cappellato et al. [13].

[11] Mondher Bouazizi and Ohtsuki Tomoaki. Participation at the author profiling shared task at pan at clef'17 http://pan.webis.de/clef17/pan17-web/author-profiling.html. 2017.

[12] 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.

[13] Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors. *CLEF 2017 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-/*, 2017. CLEF and CEUR-WS.org.

[14] Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, and Liviu P. Dinu. Including dialects and language varieties in author profiling. In Cappellato et al. [13].

[15] Heba Elfardy and Mona T Diab. Sentence level dialect identification in arabic. In *ACL (2)*, pages 456–461, 2013.

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

[17] Marc Franco-Salvador, Francisco Rangel, Paolo Rosso, Mariona Taulé, and M Antònia Martí. Language variety identification using distributed representations of words and documents. In *Experimental IR meets multilinguality, multimodality, and interaction*, pages 28–40. Springer, 2015.

[18] Marc Franco-Salvador, Nataliia Plotnikova, Neha Pawar, and Yassine Benajiba. Subword-based deep averaging networks for author profiling in social media. In Cappellato et al. [13].

[19] Barathi Ganesh and Anand Kumar. Participation at the author profiling shared task at pan at clef (http://pan.webis.de/clef17/pan17-web/author-profiling.html). 2017.

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

[24] Chu-Ren Huang and Lung-Hao Lee. Contrastive approach towards text source classification based on top-bag-of-word similarity. In *PACLIC*, pages 404–410, 2008.

[25] Andrey Ignatov, Liliya Akhtyamova, and John Cardiff. Twitter author profiling using word embeddings and logistic regression. In Cappellato et al. [13].

[26] Jamal Ahmad Khan. Author profile prediction using trend and word frequency based analysis in text. In Cappellato et al. [13].

[27] Guillaume Kheng, Léa Laporte, and Michael Granitzer. Insa lyon and uni pasau's participation at pan@clef'17: Author profiling task. In Cappellato et al. [13].

[28] Mirco Kocher and Jacques Savoy. Unine at clef 2017: Author profiling reasoning. In Cappellato et al. [13].
[29] Don Kodiyan, Florin Hardegger, Mark Cieliebak, and Stephan Neuhaus. Author profiling with bidirectional rnns using attention with grus. In Cappellato et al. [13].
[30] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. literary and linguistic computing 17(4), 2002.
[31] Nikola Ljubesic, Nives Mikelic, and Damir Boras. Language indentification: how to distinguish similar languages? In *2007 29th International Conference on Information Technology Interfaces*, pages 541–546. IEEE, 2007.
[32] 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.
[33] 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.
[34] 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 Cappellato et al. [13].
[35] Marco Lui and Paul Cook. Classifying english documents by national dialect. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 5–15. Citeseer, 2013.
[36] 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.
[37] Wolfgang Maier and Carlos Gómez-Rodrıguez. Language variety identification in spanish tweets. *LT4CloseLang 2014*, page 25, 2014.
[38] Shervin Malmasi, Mark Dras, et al. Automatic language identification for persian and dari texts. In *Proceedings of PACLING*, pages 59–64, 2015.
[39] Ilia Markov, Helena Gómez-Adorno, and Grigori Sidorov. Language- and subtask-dependent feature selection and classifier parameter tuning for author profiling. In Cappellato et al. [13].
[40] Matej Martinc, Iza Škrjanec, Katja Zupan, and Senja Pollak. Pan 2017: Author profiling - gender and language variety prediction. In Cappellato et al. [13].
[41] Yasuhide Miura, Tomoki Taniguchi, Motoki Taniguchi, and Tomoko Ohkuma. Author profiling with word+character neural attention network. In Cappellato et al. [13].
[42] Alexander Ogaltsov and Alexey Romanov. Language variety and gender classification for author profiling in pan 2017. In Cappellato et al. [13].

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

[44] 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.

[45] Adam Poulston, Zeerak Waseem, and Mark Stevenson. Using tf-idf n-gram and word embedding cluster ensembles for author profiling. In Cappellato et al. [13].

[46] 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.

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

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] Rodrigo Ribeiro-Oliveira and Rosalvo Ferreira de Oliveira-Neto. Using character n-grams and style features for gender and language variety identification. In Cappellato et al. [13].

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

[55] Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. Automatic identification of arabic language varieties and dialects in social media. *Proceedings of SocialNLP*, page 22, 2014.

[56] Nils Schaetti. Unine at clef 2017: Tf-idf and deep-learning for author profiling. In Cappellato et al. [13].

[57] 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.

[58] Sebastian Sierra, Manuel Montes y Gómez, Thamar Solorio, and Fabio A. González. Convolutional neural networks for author profiling in pan 2017. In Cappellato et al. [13].

[59] Eric S. Tellez, Sabino Miranda-Jiménez, Mario Graff, and Daniela Moctezuma. Gender and language variety identification with microtc. In Cappellato et al. [13].

[60] 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.

[61] Marcos Zampieri and Binyam Gebrekidan Gebre. Automatic identification of language varieties: the case of portuguese. In *The 11th conference on natural language processing (KONVENS)*, pages 233–237. Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI), 2012.