Fine-Grained Analysis of Language Varieties and Demographics

Francisco Rangel[1], Paolo Rosso[1], Wajdi Zaghouani[2], Anis Charfi[3]

[1]*Pattern Recognition and Human Language Technologies, Universitat Politècnica de València, Spain*
[2]*College of Humanities and Social Sciences, Hamad Bin Khalifa University, Qatar*
[3]*Information Systems Program, Carnegie Mellon University in Qatar*

Abstract

The rise of social media empowers people to interact and communicate with anyone anywhere in the world. The possibility of being anonymous avoids censorship and enables freedom of expression. Nevertheless, this anonymity might lead to cyber-security issues such as opinion spam, sexual harassment, incitement to hatred, or even terrorism propaganda. In such cases, there is a need to know more about the anonymous users and this could be useful in several domains beyond security and forensics such as marketing for example. In this paper, we focus on a fine-grained analysis of language varieties while considering also the authors' demographics. We present a Low Dimensionality Statistical Embedding (LDSE) method to represent text documents. We compared the performance of this method with the best performing teams in the Author Profiling task at PAN 2017. We obtained an average accuracy of 92.08% vs. 91.84% for the best performing team at PAN 2017. We also analyze the relationship of the language variety identification with the authors' gender. Furthermore, we applied our proposed method to a more fine-grained annotated corpus of Arabic varieties covering 22 Arab countries and obtained an overall accuracy of 88.89%. We have also investigated the effect of the authors' age and gender on the identification of the different Arabic varieties, as well as the effect of the corpus size on the performance of our method.

Keywords: language variety identification; demographics; gender; age; author profiling; cyber-security; Arabic.

1 Introduction

The rise of social media has created new ways of communication without frontiers nor censorship. Social media offers a wide range of communication possibilities to bounds never seen before. In this new (virtual) environment, millions of people share information and relate to others with their digital identity, which does not always match the real identity. Some people, in some occasions and for different reasons may want to hide their identity, omit some personal information, or highlight certain aspects to pretend being someone else. The anonymity of social

media users and the lack of knowledge about their real identity may lead to cybersecurity issues such as spreading threatening messages (Kandias *et al.* 2013), sexual harassment to minors (Inches and Crestani 2012; Bogdanova *et al.* 2014), opinion spam (Hernández-Fusilier *et al.* 2015), or even terrorism propaganda (Taylor *et al.* 2014).

Since 2017, we take part in the ARAP[1] project on author profiling for cybersecurity, which is funded by Qatar National Research Fund (QNRF) (Rosso *et al.* 2018b). One of the project aims is determining the linguistic profile of the author of a suspicious or threatening text (Russell and Miller 1977). When a suspicious message is detected, we check the veracity of the threat and discard deceptive or ironic messages. Then, if the message is considered to be a real threat, we profile the demographics of its anonymous author (Rangel and Rosso 2016a). As part of this project, we also aim at fine-grained Arabic language variety identification in combination with authors' demographics such as gender and age. To that end, we use a method to represent textual documents that considerably reduces their dimensionality, which makes it suitable for big data environments such as social media. At the same time, LDSE remains very competitive when compared to the best performing state of the art methods. To evaluate the competitiveness of our proposed method, we compare its performance with the best participating systems at the author profiling shared task of PAN 2017 (Rangel *et al.* 2017). Then, we analyse its performance using ARAP-Tweet (Zaghouani 2018a), which is a fine-grained annotated corpus covering 15 different Arabic varieties.

The rest of the paper is structured as follows. In Section 2, we report on related work. In Section 3, we present our method for representing texts and the two corpora we used. In Section 4, we present the comparative results with the best performing teams in the Author Profiling shared task at PAN 2017. Moreover, we analyse the behaviour of our proposed method with respect to the language varieties and authors' gender. In Section 5, we report on a more fine-grained Arabic language variety identification. Furthermore, we analyze several aspects related to each variety, the effect of authors' age and gender, and the impact of the corpus size on the performance. Finally, we draw some conclusions and outline future work direction in Section 6.

2 Related work

Discriminating similar languages (e.g., Malaysian vs. Indonesian) or varieties of the same language (e.g., English from UK vs. US, Spanish from Peru vs. Colombia) does not only require dealing with very similar texts at the lexical, syntactical and semantic levels, but also at the pragmatics level due to the cultural idiosyncrasies of the authors. In the last years, several researchers have addressed this task for different languages such as English (Lui and Cook 2013), Chinese (Huang and Lee 2008), Spanish (Franco-Salvador *et al.* 2015; Rangel *et al.* 2016b; Maier and

[1] `http://arap.qatar.cmu.edu`

Gómez-Rodríguez 2013), or Portuguese (Zampieri and Gebre 2012). In this context, (Zampieri and Gebre 2012) created a corpus for Portuguese by collecting 1,000 articles from the Folha de S. Paulo[2] and Dirio de Notcias[3] newsletters, respectively for Brazilian and Portugal varieties. They reported variety identification accuracies of 99.6%, 91.2%, and 99.8% with word unigrams, word bigrams and character 4-grams respectively. Also in Portuguese, (Castro *et al.* 2016) combined character 6-grams with word unigrams and bigrams allowed obtaining an accuracy of 92.71% in Twitter texts. In case of Spanish, (Maier and Gómez-Rodríguez 2013) combined language models with n-grams allowed reaching accuracies in the range of 60-70% in variety identification among Argentinian, Chilean, Colombian, Mexican, and Spanish also on Twitter texts. Similarly, the authors of (Rangel *et al.* 2016b) created the HispaBlogs[4] corpus, which covers Spanish varieties from Argentina, Chile, Mexico, Peru, and Spain. They proposed a low-dimensionality representation to represent the texts and reported accuracies of 71.1%. In another investigation with HispaBlogs, (Franco-Salvador *et al.* 2015) compared the previous representation with Skip-grams and Sentence Vectors, obtaining 72.2% and 70.8% of accuracy respectively. In case of Chinese, (Xu *et al.* 2016) combined general features such as character and word n-grams with PMI-based and word alignment-based features to approach the task of identifying among varieties of Mandarin Chinese for the Greater China Region: Mainland China, Hong Kong, Taiwan, Macao, Malaysia, and Singapore. They reported accuracies up to 90.91%

The interest in language variety identification is also reflected by the number of tasks that were organised in the last years:

- Defi Fouille de Textes (DEFT) shared task (Grouin *et al.* 2011) on language variety identification of French texts was organised in 2010.
- LT4CloseLang workshop on Language Technology for Closely Related Languages and Language Variants shared task was organised at EMNLP 2014 (Agić *et al.* 2014).
- VarDial Workshop (Zampieri *et al.* 2014) on applying NLP Tools to Similar Languages, Varieties and Dialects was organised in 2014 at the International Conference on Computational Linguistics (COLING). The workshop focused on 13 language varieties: Bosnian, Croatian, Serbian; Indonesian, Malay; Czech, Slovak; Brazilian Portuguese, European Portuguese; Peninsular Spanish, Argentinian Spanish; and American English, British English. The best performance was obtained with a two-step approach with word and char n-grams as features. The language group was predicted with a probabilistic model and then SVM was used to discriminate within each group.
- LT4Vardial joint workshop on Language Technology for Closely Related Languages, Varieties, and Dialects (Zampieri *et al.* 2015) was organised in 2015 at RANLP. It focused on 13 languages grouped as follows: Bulgarian, Macedonian; Bosnian, Croatian, Serbian; Czech, Slovak; Malay, Indonesian; Brazilian,

[2] `http://www.folha.uol.com.br`
[3] `http://www.dn.pt`
[4] `https://github.com/autoritas/RD-Lab/tree/master/data/HispaBlogs`

European Portuguese; Argentinian, Peninsular Spanish; and a group with a variety of other languages. The best performing team used an ensemble of SVM classifiers and character n-grams.

- Vardial workshop on NLP for Similar Languages, Varieties, and Dialects (Malmasi *et al.* 2016) was organised in 2016 at COLING, with the following two subtasks: (i) a more realistic task with the removal of very easy to discriminate languages such as Czech vs. Slovak and Bulgarian vs. Macedonian, and including new varieties such as Hexagonal vs. Canadian French; and (ii) a new subtask on discriminating Arabic dialects in speech transcripts with Modern Standard Arabic and four Arabic dialects (Egyptian, Gulf, Levantine, and North African). The best result was obtained with SVM ensembles by the same team who ranked first in DSL 2015.
- Vardial Evaluation Campaign (Zampieri *et al.* 2017) was organised at EACL 2017, with four shared tasks: (i) Discriminating Between Similar Languages; (ii) Arabic Dialect Identification; (iii) German Dialect Identification; (iv) Cross-lingual Dependency Parsing. The best result was obtained with a Kernel Discriminant Analysis classifier trained on a combination of n-grams based kernels such as the sum of a blended presence bits kernel and a blended intersection kernel, together with a kernel based on LRD with 3 to 7 characters, and a quadratic RBF kernel based on i-vectors.
- Author Profiling at PAN 2017 (Rangel *et al.* 2017) focused on language variety identification in combination with gender identification. The task addressed four languages: (i) English (Australia, Canada, Great Britain, Ireland, New Zealand, United States); (ii) Spanish (Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela); (iii) Portuguese (Brazil, Portugal); and (iv) Arabic (Egypt, Gulf, Levantine, Maghreb). The best results were obtained with traditional machine learning approaches (SVM, logistic regression) and combinations of n-grams and hand-crafted features such as the occurrence of emojis, sentiments, or lists of words per variety.

Along the same lines, we witnessed recently an increasing interest in Arabic varieties identification as shown by the high number of teams that participated in the Arabic subtask of the third (Malmasi *et al.* 2016) DSL track (18 teams) and in the Arabic Dialect Identification (ADI) shared task (Zampieri *et al.* 2017), as well as in the Arabic subtask of the Author Profiling shared task (Rangel *et al.* 2017) at PAN 2017 (20 teams). However, as (Rosso *et al.* 2018a) highlighted, there is still a lack of resources for Arabic and research works that are specific to that language. Some of the few works are mentioned in the following. In (Zaidan and Callison-Burch 2014), Zaidan et al. used a smoothed word unigram model and reported respectively 87.2%, 83.3% and 87.9% of accuracies for Levantine, Gulf and Egyptian varieties. In (Sadat *et al.* 2014), the authors achieved 98% of accuracy discriminating among Egyptian, Iraqi, Gulf, Maghreb, Levantine, and Sudan with n-grams. In (Elfardy and Diab 2013), combined content and style-based features allowed to obtain 85.5% of accuracy when discriminating between Egyptian and Modern Standard Arabic.

Nonetheless, to the best of our knowledge, the first time that the language variety identification was combined with demographic traits such as authors' gender was at PAN'17, and there are no other investigations that focus on the combined analysis of both aspects (language variety and demographics). Furthermore, in case of Arabic most research focused on coarse-grained groups of regional language varieties (e.g., Levantine, Maghreb, Gulf, etc.) and did not work on fine-grained analysis (i.e., at the country level).

3 Evaluation framework

In this section, we present the Low Dimensionality Statistical Embedding method to represent documents, as well as the two corpora we used to evaluate its performance[5].

3.1 Low Dimensionality Statistical Embedding

Low Dimensionality Statistical Embedding (LDSE) is the generalisation of the Low Dimensionality Representation (LDR) method (Rangel *et al.* 2016b) where skewness, kurtosis, and moments (Bowman and Shenton 1985) are used to measure the distribution of weights for each class. The intuition behind both methods is that, in an annotated corpus, the probability of each term to belong to each of the classes should be different. If we use weights to represent such probability, we may assume that the distribution of weights for a given document should be closer to the weights of its corresponding class.

We obtain the *tf-idf* (Salton and Buckley 1988) matrix (Equation 1) for the terms of the documents D in the training set. Each row represents a document d_i and each column represents a term t_j belonging to the vocabulary T. Each w_{ij} represents the *tf-idf* weight for the term t_j in the document d_i. The last column $\delta(d_i)$ represents the assigned class c from the set of all classes C to the document d_i.

$$\begin{bmatrix} w_{11} & w_{12} & ... & w_{1m} & \delta(d_1) \\ w_{21} & w_{22} & ... & w_{2m} & \delta(d_2) \\ ... & ... & ... & ... & \\ w_{n1} & w_{n2} & ... & w_{nm} & \delta(d_n) \end{bmatrix}, \tag{1}$$

As formalised in Equation 2, for each term t and each class c, we define the term weight $W(t, c)$ as the ratio between the weights of the documents belonging to the class c and the sum of all weights for that term.

[5] We use accuracy to evaluate the systems as: *i)* the corpora are completely balanced; *ii)* in case of PAN, we can compare our obtained results with the official ones. Since accuracy is the proportion of properly classified instances, we apply the two population proportions hypothesis test to determine the significance of the results (McNemar 1947).

$$W(t,c) = \frac{\sum_{d \in D/c=\delta(d)} w_{dt}}{\sum_{d \in D} w_{dt}}, \forall d \in D, c \in C \quad (2)$$

A document d is represented as shown in Equation 3, with as many dimensions as the number of terms in the document multiplied by the number of classes.

$$\begin{aligned} d = W(t,c) = \\ \{W(t_1,c_1), W(t_2,c_1), ..., W(t_t,c_1), \\ W(t_1,c_2), W(t_2,c_2), ..., W(t_t,c_2), \\ ..., \\ W(t_1,c_c), W(t_2,c_c), ..., W(t_t,c_c)\} \\ \sim \forall\ t \in T, c \in C \end{aligned} \quad (3)$$

In order to reduce the dimensionality of the representation, we obtain descriptive statistics from the previous distribution of weights. (Heitele 1975) pointed out three fundamental concepts regarding random variables[6]: their distribution, mean and variability. Moments are based on a generalisation of the average. Hence, they are generic indicators of the distribution. They represent the arithmetic mean of a specified integer power of the deviation of the variable from the mean. In this sense, two distributions are equal if all their infinite moments coincide. Thus, we can assume that the more similar both distributions are, the more similar their moments are. For the distribution of weights for each class c we obtain the following measures SE (Statistical Embedding) shown in Equation 4: minimum, maximum, average, median, first and third quartiles (Q), (Gini 1971) indexes (G) to measure the distribution skewness and kurtosis, and the first ten moments (M). Based on that, documents are represented using Equation 5.

$$\begin{aligned} SE(W) = \{&min(W), max(W), avg(W), median(W), \\ &Q_1(W), Q_3(W), G_1(W), G_2(W), M_{2..10}(W)\} \end{aligned} \quad (4)$$

$$d = SE(W(t,c)) \sim \forall\ t \in T, c \in C \quad (5)$$

To better illustrate the previous formulas and their practical application, we used the LDSE method to represent the documents of a corpus annotated with two classes. This corpus is the Portuguese subset of the PAN-AP17, which will be explained later and for which the average feature avg(W) is plotted in Figure 1. This figure confirms that both classes can be easily separated.

We experimented with several machine learning algorithms (Bayesian, Logistic, Neural Networks, Support Vector Machines, Trees and Rules-based, Lazy, and

[6] Despite the fact that we cannot assume randomness on the distribution of weights, somehow the presented descriptive statistics can summarise their distribution.

Meta-classifiers) implemented in Weka[7]. After that, we selected the best performing ones on the training data in each case.

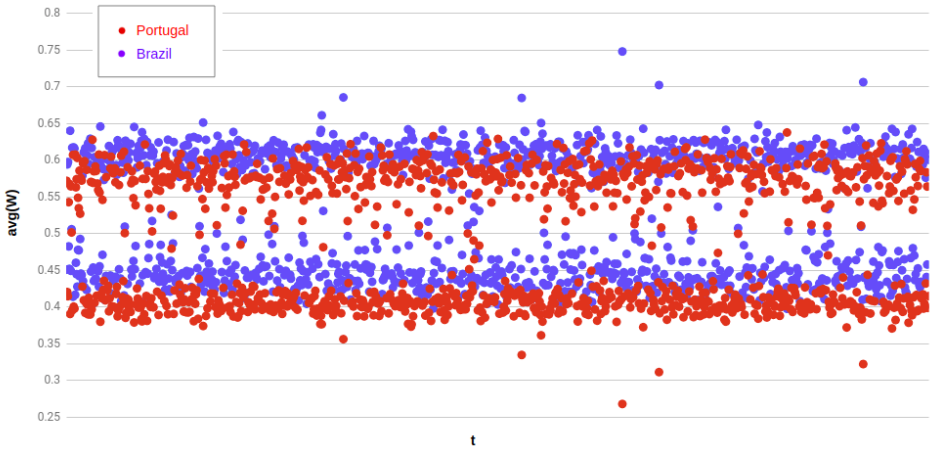

Fig. 1. Portuguese subset of the PAN-AP'17 represented with the avg(W) feature of LDSE. The X-axis represents each of the terms in the corpus; The Y-axis represents the average weight for each term in Brazilian (blue) or Portuguese (red) varieties.

3.2 Corpora

In this section, we describe the corpora used in this research work. First, we describe the PAN-AP'17 corpus which covers four languages and their varieties. This corpus allowed us to demonstrate the suitability of LDSE to address language variety identification, taking into account also the authors gender. Then, we describe the ARAP-Tweet corpus (Zaghouani 2018a) which allows us to evaluate the use of the LDSE method for more fine-grained identification of Arabic varieties taking into account the authors' age and gender.

3.2.1 PAN-AP'17

PAN Lab[8] at CLEF (Conferences and Labs of the Evaluation Forum)[9] focuses on different forensics linguistics tasks: author identification (Kestemont *et al.* 2018), profiling (Rangel *et al.* 2018), and obfuscation (Hagen *et al.* 2018). Given a certain document, the aims are to infer who is the author that wrote it as well as the authors demographic traits. Obfuscation is the opposite task to author identification. It aims at making the identification of authors based on their writting style impossible. PAN provides an opportunity for the research community to validate and compare the

[7] `https://www.cs.waikato.ac.nz/ml/weka/`
[8] `https://pan.webis.de/`
[9] `http://www.clef-initiative.eu`

state-of-the-art methods and technologies for the three forensics linguistics tasks mentioned above.

The focus of the 2017 Author Profiling shared task was on gender and language variety identification in Twitter. The PAN-AP'17 corpus includes four languages: Arabic[10], English, Portuguese and Spanish. For each language several varieties were considered as shown in Table 1. For each variety, tweets geolocated in the capital cities (or the most populated cities) where this language variety is used were collected. Unique users were selected and annotated with their corresponding variety. A dictionary with proper nouns was used to annotate the users' gender. Moreover, we manually inspected their profile photo to improve the annotation quality. Finally, for each user one hundred tweets were collected from her/his timeline. The corpus was divided into training/test following a 60/40 proportion, with 300 authors for training and 200 authors for test per gender and variety. More information on this corpus is available in the shared task overview paper (Rangel *et al.* 2017).

Table 1. *PAN-AP'17 corpus, covering four languages with their corresponding varieties and the cities selected as representative of such varieties.*

Language	Variety	City
Arabic	Egypt	Cairo
	Gulf	Abu Dhabi, Doha, Kuwait, Manama, Mascate, Riyadh, Sana'a
	Levantine	Amman, Beirut, Damascus, Jerusalem
	Maghreb	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

[10] In case of Arabic, the selection of these varieties corresponds to previous works (Sadat *et al.* 2014). Iraqi was selected and then discarded due to the lack of enough tweets.

3.2.2 ARAP-Tweet

ARAP-Tweet is a corpus that was developed at Carnegie Mellon University Qatar (Zaghouani 2018a) in the context of the ARAP project. The total number of tweets in this corpus is above 2 millions (exactly 2,032,539) and the total number of words is above 18 millions (exactly 18,582,436). Across all dialectal varieties of this corpus, the average number of tweets per user is 684 and the average number of words per tweet is 9.

Arabic dialects have been generally classified by regions such as in (Habash 2010), who classified the Arabic major dialects into North African, Levantine, Egyptian and Gulf. Similar dialectical varieties were also used at PAN following (Sadat *et al.* 2014). However, dialect variation within regions could be significant. For example, the Tunisian dialect is different from the Moroccan dialect even though they belong to the same North African/Maghreb region. Therefore, fine-grained annotated Arabic language resources are required. ARAP-Tweet is a corpus that provides fine-grained dialectal Arabic tweets annotated with age and gender information. It contains 15 dialectical varieties corresponding to 22 countries of the Arab world. For each variety, a total of 102 authors (78 for training, 24 for test) were annotated with age and gender, maintaining balance for both variables. Three age groups are distinguished: Under 25, Between 25 and 34, and Above 35. The included varieties, as well as the regions they belong to, are shown in Table 2. Further information on this corpus is available in (Zaghouani 2018a; Zaghouani 2018b).

Table 2. *ARAP-Tweet corpus: language varieties and regions.*

Language Variety	Region (Sadat *et al.*, 2014)
Algeria	Maghrebi
Egypt	Egyptian
Iraq	Iraqi
Kuwait	Gulf
Lebanon Syria	Levantine
Libya	Maghrebi
Morocco	Maghrebi
Oman	Gulf
Palestine Jordan	Levantine
Qatar	Gulf
Saudi Arabia	Gulf
Sudan	Other
Tunisia	Maghrebi
United Arab Emirates (UAE)	Gulf
Yemen	Gulf

4 Language variety identification at PAN'17

In this section, we compare LDSE with the best performing teams of the 22 participants in the Author Profiling shared task at PAN 2017. We also analyse the obtained results from two perspectives: the confusion among varieties of the same

language and the effect of the gender on language variety identification. Finally, we discuss the suitability of the LDSE method to the task of language variety identification.

4.1 Classification results

Figure 2 shows the results obtained by the three best performing teams at PAN 2017 together with the results we obtained with LDSE[11]. Results are shown for the four languages as well as the average among them. At PAN, the best accuracy results in Arabic and Spanish were achieved by (Basile *et al.* 2017), who obtained 83.13% and 96.21% respectively. They also obtained the best overall result in the shared task (91.84%). In case of English and Portuguese the best accuracy was obtained by (Tellez *et al.* 2017), with 90.04% and 98.5% respectively. Overall, they had the second best result in the task (91.71%). Basile's team approached the task with combinations of character, tf-idf word n-grams, and SVM. Similarly, Tellez's team used SVM with combinations of bag-of-words. The third best performing team was (Martinc 2017), who used logistic regression with combinations of character, word, POS n-grams, emojis, sentiments, character flooding, and lists of words per variety, achieving an average accuracy of 90.85%. It is worth mentioning that also deep learning approaches (e.g., Recurrent Neural Networks, Convolutional Neural Networks, as well as word and character embeddings) were used by other participants but they did not lead to the best results.

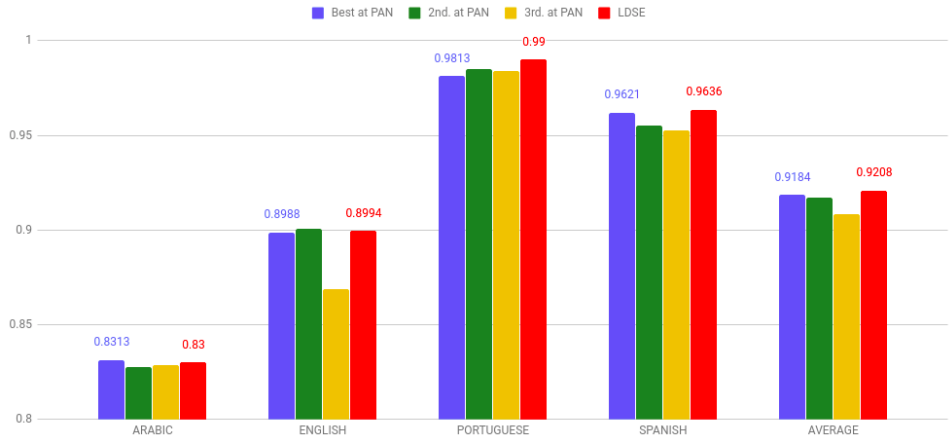

Fig. 2. Comparative results of the three best performing teams in the Author Profiling shared task at PAN 2017 vs. LDSE. The best performing team (Basile *et al.* 2017) obtained the highest result in Arabic and Spanish. The second best performing team (Tellez *et al.* 2017) obtained the highest result in English and Portuguese.

[11] We have used the following machine learning methods: *i)* BayesNet for Arabic; *ii)* SVM for Spanish; and *iii)* Random Forest for English and Portuguese.

Table 3. *Significance (p-values) when comparing LDSE results with the three best performing teams in the Author Profiling shared task at PAN 2017 (*0.05; **0.01).*

	Arabic	English	Portuguese	Spanish	Average
Best at PAN	-0.0980	0.0690	4.4631	0.2968	0.5441
2nd. at PAN	0.1877	-0.1154	0.9001	1.5564	0.8357
3rd. at PAN	0.0902	3.3115*	1.0906	2.0717**	2.7137*

In Figure 2, the results obtained by LDSE are also shown. The figure shows that LDSE achieves the best results for Portuguese (99% vs. 98.5%) and Spanish (96.36% vs. 96.21%), while it achieves the second best results for Arabic (83% vs. 83.13%) and English (89.94% vs. 90.04%). Overall, LDSE has the best performance with an average accuracy of 92.08% vs. the second best performance of 91.84%. As shown in Table 3, there is no statistical significance between the best results at PAN and the ones obtained by LDSE, which confirms its competitiveness with the state-of-the-art approaches.

4.2 Confusion among varieties

The error among varieties of the same language is analysed using confusion matrices[12] as shown in Figures 3, 4, 5, and 6 respectively for Arabic, English, Portuguese and Spanish.

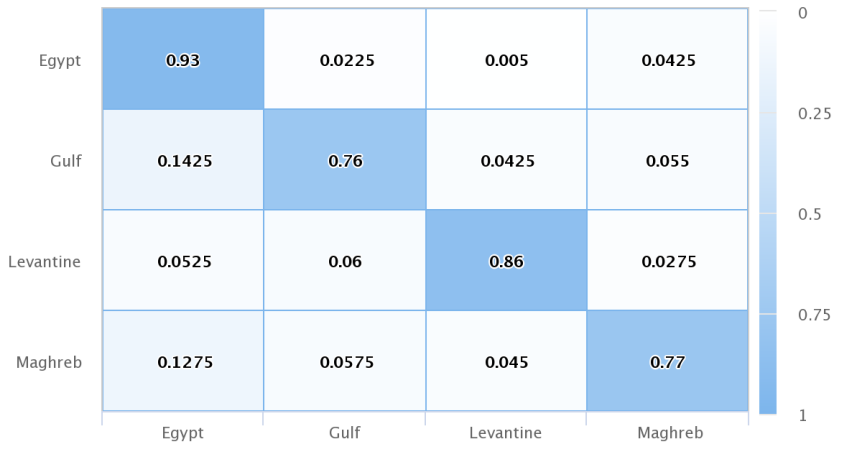

Fig. 3. Confusion matrix for Arabic varieties with LDSE on the PAN-AP'17 corpus.

[12] Matrices show the percentage (in the range 0..1) of instances classified in each variety (per row) that actually belongs to the variety in the columns.

As shown in Figure 3, the maximum confusion in Arabic varieties is from Gulf to Egypt (14.25%), followed by Maghreb to Egypt (12.75%) whereas the lowest confusion is from Egypt to Levantine (0.5%). The rest of the errors are between 2.25% (from Egypt to Gulf) and 6% (from Levantine to Gulf). The highest accuracy was obtained for the identification of Egyptian Arabic (93%). Together with the lowest confusion seen previously, these results show that this variety is the less difficult to be identified. Conversely, the Gulf and Maghreb varieties are the most difficult ones to identify, with accuracies of 76% and 77% respectively, and with the highest confusions to other varieties. Finally, the results obtained for the Levantine variety are higher than the average (86% over 83%). These results are similar to the ones obtained by the PAN participants, where both the Egyptian and Levantine Arabic varieties were the less difficult to identify.

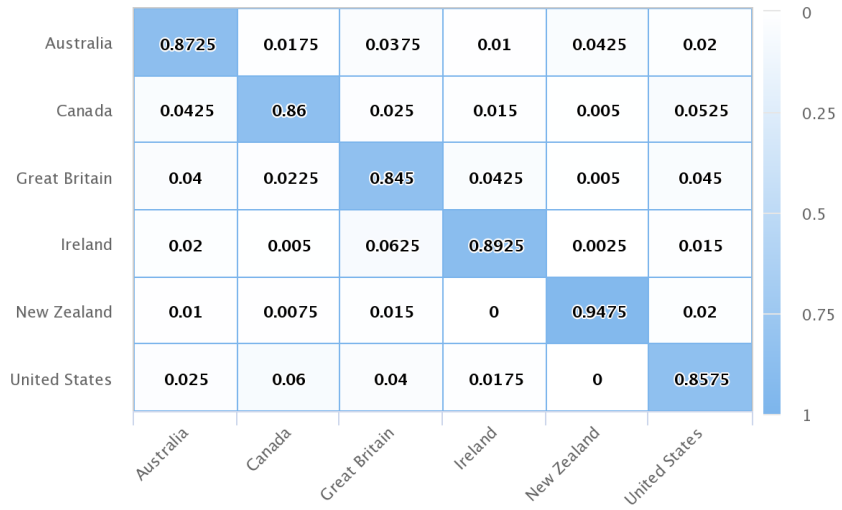

Fig. 4. Confusion matrix for English varieties for LDSE on the PAN-AP'17 corpus.

Figure 4 shows the LDSE confusion matrix among English varieties. The highest confusion is from Ireland to Great Britain (6.25%), United States to Canada (6%), Canada to United States (5.25%), and Great Britain to United States (4.5%). Some of these errors correspond to varieties geographically close or that even share geographical borders. The other errors are lower than 4.5%, with almost no error in cases such as New Zealand to Canada (0.75%), Canada or Great Britain to New Zealand (0.5%), Ireland to Canada (0.5%), Ireland to New Zealand (0.25%), New Zealand to Ireland (0%) and United States to New Zealand (0%). Considering the previous insights, together with the highest accuracy obtained (94.75%), we conclude that the New Zealand variety is the less difficult English variety to identify. The second less difficult English variety is Irish English (89.25%), and all the rest range between this maximum value (of 89.25%), and the minimum value obtained

for Great Britain (84.5%). Similarly to what was observed already at PAN, we conclude that the geographically closer two English varieties are, the higher the confusion between them is.

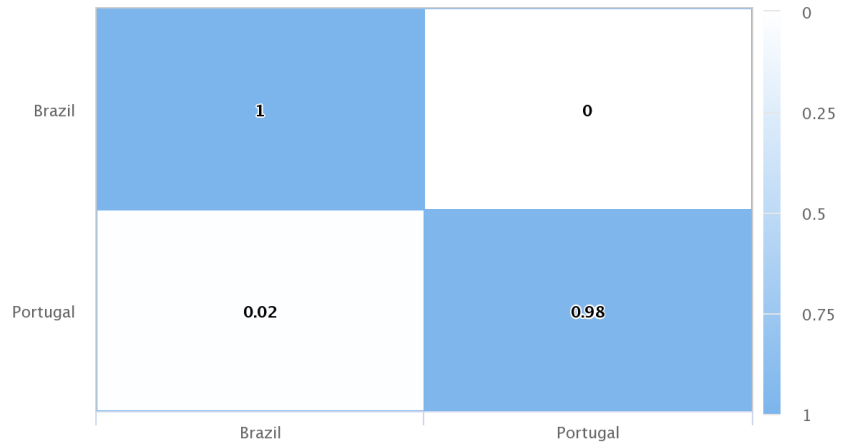

Fig. 5. Confusion matrix for Portuguese varieties for LDSE on the PAN-AP'17 corpus.

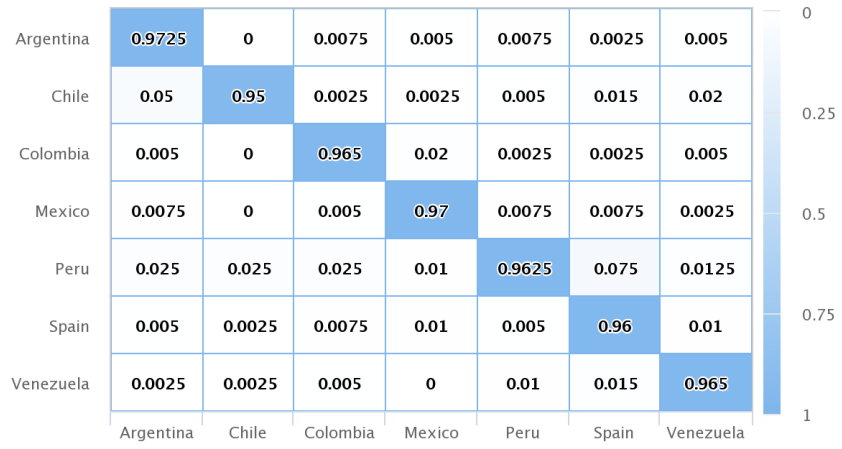

Fig. 6. Confusion matrix for Spanish varieties for LDSE on the PAN-AP'17 corpus.

As shown in Figure 5, the results for Portuguese are very high and almost without

errors, which is in line with the results achieved by the PAN shared task participants. There is no confusion from Brazil to Portugal varieties, and only 2% of the Portuguese variety is confused with the Brazilian one. This gives an accuracy of 100% for identifying Brazilian Portuguese, which is the less difficult Portuguese variety to be identified. The accuracy is 98% for the Portuguese variety of Portugal.

In case of Spanish, the confusion matrix among varieties is shown in Figure 6. It can be observed that all the Spanish varieties have similar results, ranging from 95% to 97.25%, with no significant difference among them. The highest error is from Peru to Spain (7.5%), Chile to Argentina (5%), Peru to Argentina, Chile and Colombia (2.5%), and the rest are lower to 2%. Except in the case of Peru and Spain, we can conclude again that the geographical proximity of varieties may affect the confusion between them.

Finally, in Table 4 we summarise the differences between the lowest and highest accuracy obtained for each language both for the best participant at PAN and for LDSE. The last column shows the difference between PAN and LDSE. In case of English and Spanish LDSE is significantly more stable than the systems at PAN. This is also true for Portuguese but without statistical significance. In case of Arabic, LDSE is significantly more variable. However, we can argue in favour of this variability due to the very high accuracy obtained for the Egyptian variety (93%, about 10% higher than the best performing team at PAN).

Table 4. *Difference between highest and lowest accuracies per variety language, both for the best participant at PAN in that language and LDSE. The last column shows the difference between them (* indicates a significant difference).*

Language	PAN	LDSE	Diff.
Arabic	0.0942	0.1700	-0.0758*
English	0.1656	0.1025	0.0631*
Portuguese	0.0352	0.0200	0.0152
Spanish	0.1083	0.0225	0.0858*

4.3 The impact of the gender on the language variety identification

In this section, we compare the systems at PAN to LDSE with respect to the impact of gender on the language variety identification. In Table 5, we compare the LDSE results to the average results of the systems at PAN, as reported in (Rangel *et al.* 2017). In Table 6, we compare LDSE to the best performing system per language at PAN 2017[13]. Both tables show that it is more difficult to properly identify the variety in case of males except for Spanish. We also observe that at PAN, these differences are significant in case of Arabic and Portuguese. Especially in the case of Arabic, the difference decreases from 7.06% to 2.50%.

[13] (Basile *et al.* 2017) in Arabic and Spanish; (Tellez *et al.* 2017) in English and Portuguese.

Table 5. *Language variety identification accuracy per gender (* indicates a significant difference) comparing LDSE to the average of all the systems at PAN.*

	PAN			LDSE			
Language	Female	Male	Diff.	Female	Male	Diff.	Diff.
Arabic	0.7909	0.7203	0.0706*	0.8425	0.8175	0.0250	0.0456*
English	0.7190	0.7168	0.0022	0.8875	0.8717	0.0158	-0.0136*
Portuguese	0.9829	0.9633	0.0196*	0.9950	0.9850	0.0100	0.0096
Spanish	0.8680	0.8733	-0.0053	0.9657	0.9614	0.0043	0.0010

When comparing LDSE to the best performing system per language at PAN (in Table 6), we can see that the difference decreases in the case of Arabic, English and Spanish whereas it remains the same in the case of Portuguese. It is noteworthy that in the case of Arabic and Spanish, the decrease is statistically significant, from 5.75% to 2.50% and from 1.00% to 0.43% for both languages.

Table 6. *Language variety identification accuracy per gender and language (* indicates a significant difference) comparing LDSE to the best performing system at PAN.*

	PAN			LDSE			
Language	Female	Male	Diff.	Female	Male	Diff.	Diff.
Arabic	0.8600	0.8025	0.0575*	0.8425	0.8175	0.0250	0.0325*
English	0.9092	0.8917	0.0175	0.8875	0.8717	0.0158	0.0017
Portuguese	0.9900	0.9800	0.0100	0.9950	0.9850	0.0100	0.0000
Spanish	0.9671	0.9571	0.0100	0.9657	0.9614	0.0043	0.0057*

In Figure 7, the errors per gender for each variety are shown in detail. In case of Arabic, we can observe that the maximum error occurs with the Gulf variety for males (31%), followed by the Maghreb variety for females (28.5%). This coincides with the analysis of the confusion matrix where we concluded that the Gulf and Maghreb varieties are the most difficult to identify. Errors per gender for both Egypt and Levantine varieties are more well-balanced, even though it is remarkable that in case of Egypt, females are a little bit more difficult to be identified (1%). In case of English there is the same number of varieties with a higher number of errors in one gender or the other. For example, in the case of Australia, New Zealand and United States, there are more errors in case of females, whereas the contrary occurs with Canada, Great Britain and Ireland. Finally, the highest difference occurs with Canada (8%). With respect to Spanish, except for Colombia and Mexico, there are more errors for males. In Spanish, the differences are smaller (2%) and the performance per gender is more balanced than in English and Arabic. In the case of Portuguese, all errors occurred with the variety from Portugal, with 3/4 of the errors belonging to females.

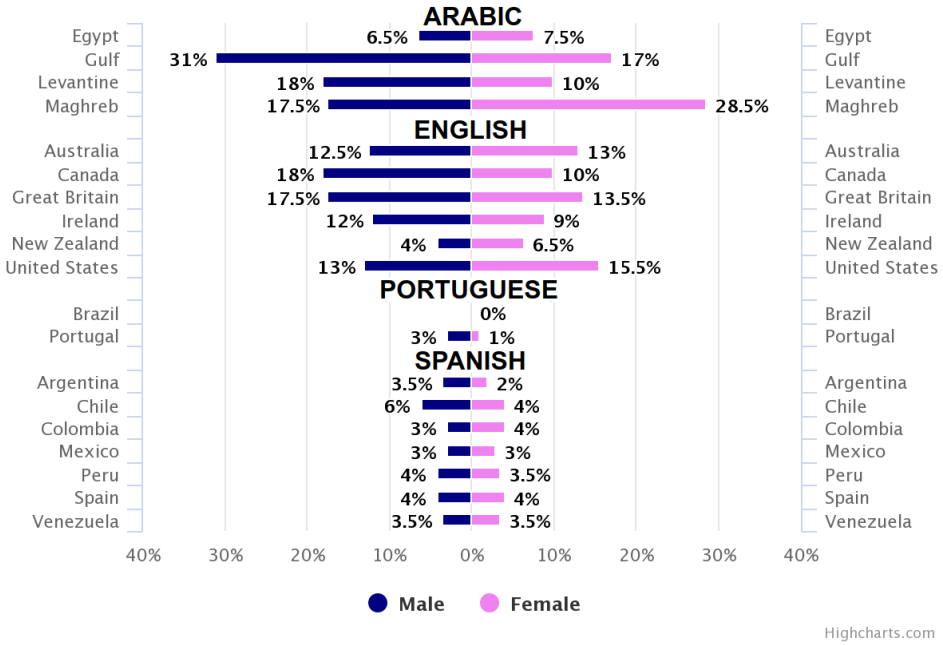

Fig. 7. Percentage of errors per gender for each language variety (PAN-AP'17 corpus).

5 Fine-grained Arabic language variety identification

We are interested in investigating further language variety identification in Arabic due to the low results obtained in comparison with the other languages (cf. Figure 2), the lack of resources for this language, and its importance for cybersecurity (Rosso *et al.* 2018a). In this section, we use the ARAP-Tweet corpus to evaluate further the performance of LDSE for the fine-grained identification of Arabic language varieties. We also study the confusion among the different Arabic varieties, together with the impact of authors' age and gender.

5.1 Classification results

Figure 8 shows the LDSE results when using the ARAP-Tweet corpus. We experimented with five machine learning algorithms: BayesNet, Multilayer Perceptron, Simple Logistics, SVM, and Random Forest.

We obtained the best accuracy result with Multilayer Perceptron (88.89%), followed by Logistic Regression and Support Vector Machines (87.5%), Random Forest (86.94%) and Bayesian Networks (86.11%). However, these differences are not statistically significant. In the next sections, we will use Multilayer Perceptron. It is worth to mention that for this experiment we selected only 100 tweets per author in order to maintain a comparable scenario to PAN, at which LDSE achieved 83% of accuracy.

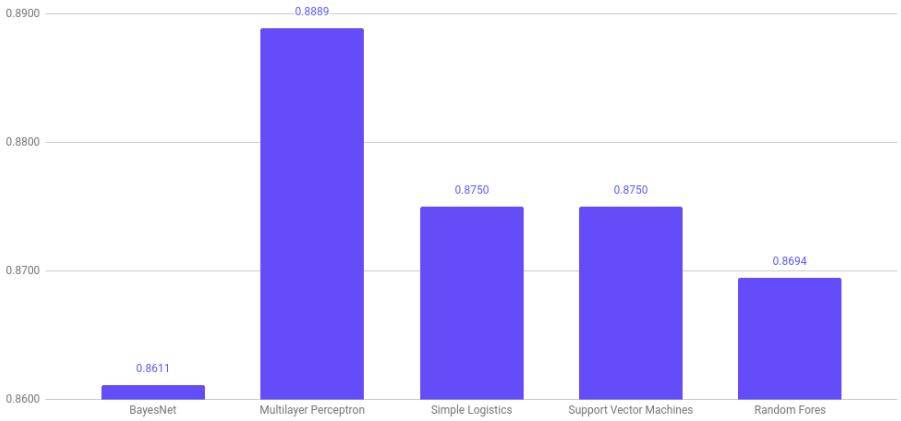

Fig. 8. Accuracy obtained by LDSE with five different machine learning classifiers on the ARAP-Tweet corpus.

5.2 Confusion among varieties

If we analyse the confusion matrix among the varieties of Figure 9, we can see that most errors occur with the Saudi variety (63% of accuracy), followed by the Qatari variety (71% of accuracy). The average accuracy was 88.89%. The following Arabic varieties were the less difficult to distinguish: Egypt (100%), Libya (100%), Morocco (100%), Sudan (100%), Iraq (96%), Lebanon Syria (92%), Palestine Jordan (92%), Tunisia (92%) and Yemen (92%). Together with Saudi Arabia and Qatar, the most difficult Arabic varieties to identify are those of Kuwait (83%), Oman (83%), and UAE (83%).

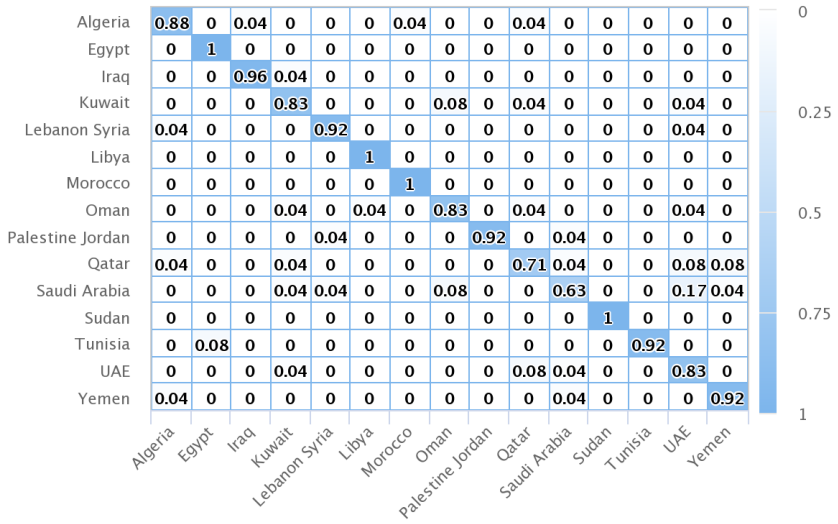

Fig. 9. Confusion matrix for Arabic varieties for LDSE on the ARAP-Tweet corpus.

The highest error occurs from Saudi Arabia to UAE (17%), varieties from two neighbouring countries. Similarly, most errors occur within the same region. For example, within Gulf there is confusion at classifying from Qatar to UAE (8%), Saudi Arabia (4%) or Yemen (8%), as well as from Kuwait to Oman (8%), Qatar (4%) and UAE (4%). Similarly, within the Levant region, there is confusion from Palestine Jordan to Lebanon Syria (4%), or to close countries albeit these are in another region. This is the case for example from Palestine Jordan to Saudi Arabia (4%) or Saudi Arabia to Lebanon Syria (4%). Similarly to PAN, the highest confusion occurs within the Arabic variety of the Gulf region whereas the highest accuracy was obtained for the identification of the Egyptian variety (100%).

5.3 The impact of the age and gender on the language variety identification

In this section, we analyse the impact of the authors' age and gender on Arabic language variety identification. Figure 10 and Table 7 show the distribution of errors depending on the authors' age and gender.

Table 7. *Distribution of the errors depending on the authors' age and gender (ARAP-Tweet corpus).*

Gender	Under 25	Between 25-34	Above 35	Total
Female	0.275	0.225	0.125	0.625
Male	0.125	0.075	0.175	0.375
Total	0.400	0.300	0.300	

We observe that the percentage of errors in case of female authors (62.5%) is much higher than in case of males (37.5%). Concretely, there is a difference of 25%. This is also true in case of the age classes *Under 25* and *Between 25-34*, where the difference is 15%. However, the opposite occurs for the age class *Above 35*, where the errors in case of males are 5% higher than in case of females. In case of females, the highest error occurs with the age class *Under 25* (27.5%), and the lowest with the age class *Above 35* (12.5%), with a significant difference of 15%. Conversely, the highest error in case of males occurs with the age class *Above 35* (17.5%) whereas the lowest one occurs with the age class *Between 25-34* age class (7.5%), with a highly significant difference of 10%. Taking into account only age ranges, the highest error is in case of the class *Under 25* (40%), with a significant difference of 10% over the other two classes. We can conclude that Arabic varieties included in ARAP-Tweet are less difficult to identify when the author is male (37.5% of error) or belongs to the age classes *Between 25-34* and *Above 35* (30% of error), and especially when the author is male in the age class *Between 25-34* (7.5% of error).

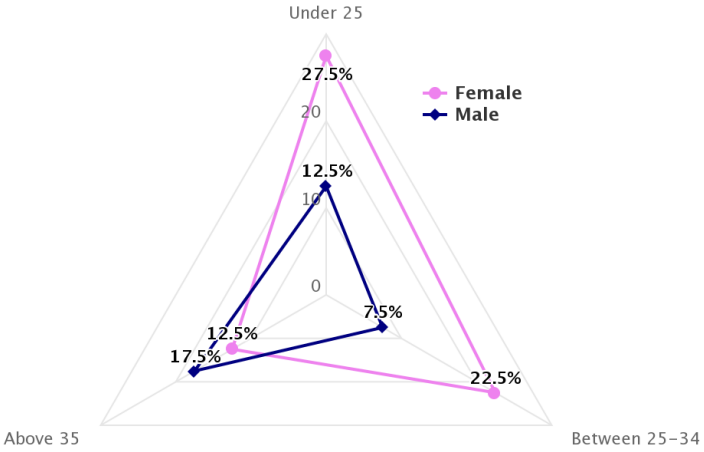

Fig. 10. Distribution of the errors depending on the authors' age and gender (ARAP-Tweet corpus).

It is noteworthy that the obtained distribution of errors per gender for this corpus is the contrary to the error distribution obtained for the PAN-AP'17 corpus. In that latter corpus, the proportion of errors between females and males was approximately 46% vs. 54%. This significant difference in error distribution can be explained by the different methodologies followed to build the two corpora. In case of ARAP-Tweet, the corpus was collected from Twitter and then perfectly balanced with respect to gender and age classes, whereas in case of PAN-AP'17, the retrieved tweets followed a real scenario distribution with respect to age groups (e.g., it included more people above 35 than under 25). Furthermore, in case of the PAN-AP'17, the collected Twitter authors had their geolocalisation activated. Probably, this option depends on the users age (e.g., younger people could be more conscious about their privacy and therefore deactivate this option more often).

In Figure 11 the error per age and gender is shown for each Arabic variety (only varieties with errors). The highest error occurs with males in the class *Above 35* in the case of Tunisia (6.98%), followed by Kuwait (4.65%) also for the same age group and gender, and Qatar (4.65%) for males in the age class *Between 25-34*. The remaining errors with males occur mainly in the age classes *Under 25* and *Above 35*, with a frequency of 2.33% each. In the case of females, the highest errors for Oman variety (6.98%) occur in the classes *Under 25* and *Between 25-34*. For Qatar, the highest errors (6.98%) occur for the class *Under 25*. For Saudi Arabia, the highest errors occur in the case of the classes *Between 25-34* (6.98%) and *Under 25* (4.65%). We should highlight the case of Kuwait and UAE that has an average error of 17% since there is an age class with no errors: *Under 25* in case of Kuwait and *Above 35* in the case of UAE. Finally, it is worth to mention that in most

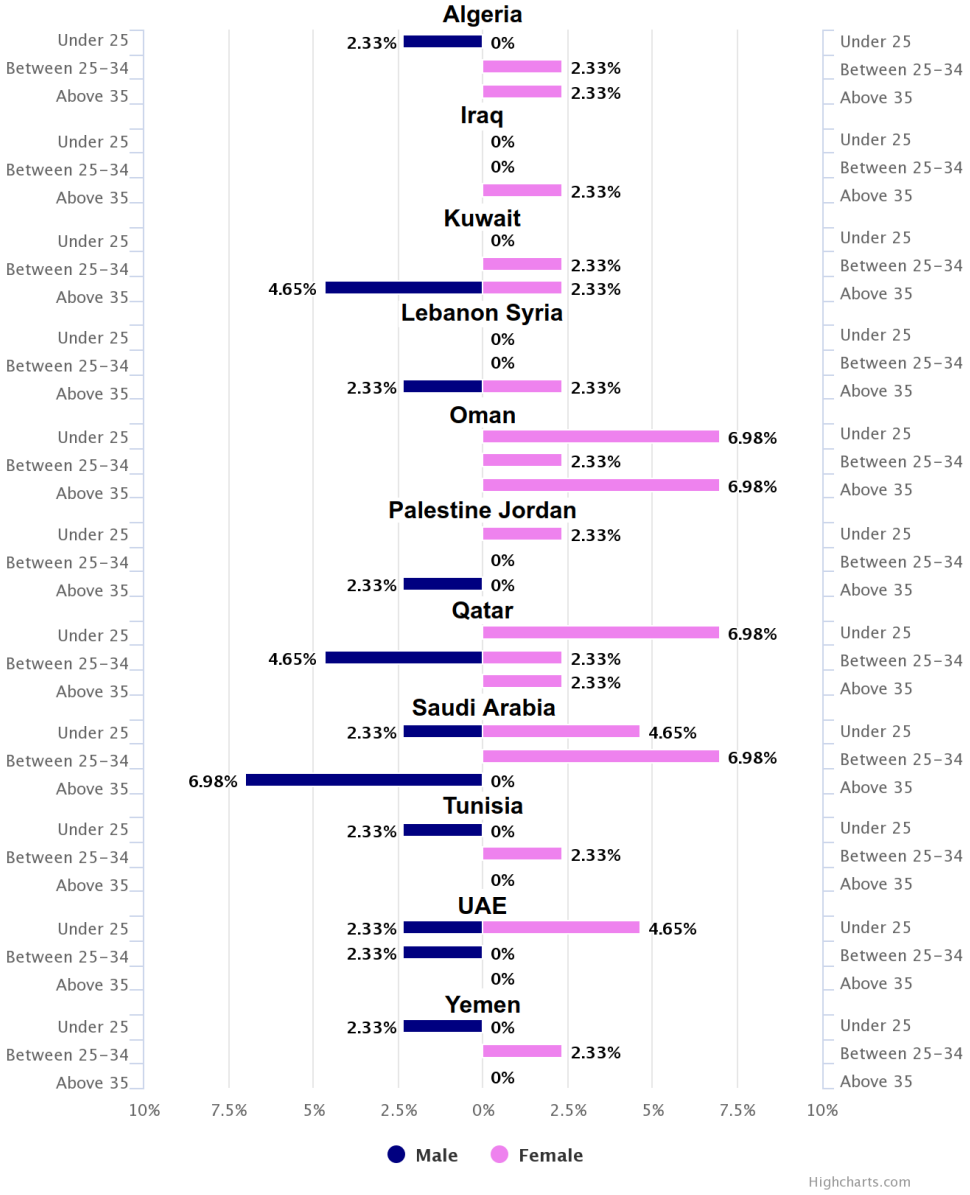

Fig. 11. Percentage of errors per age and gender for each language variety (ARAP-Tweet corpus).

Arabic varieties there are no errors for males in the class *Between 25-34* (except Qatar with 4.65% and UAE with 2.33%).

5.4 The effect of the corpus size

Since the ARAP-Tweet corpus contains a variable number of tweets per author, we analysed the effect of this number on the variety identification task using the same machine learning algorithms as described previously. In Figure 12, we observe that the accuracy of all classifiers improves when the number of tweets increases except in the case of Simple Logistics, whose behaviour becomes erratic from 700 tweets. The average accuracy increases from 87.38% to 94.79% (with the exception of Logistic Regression from 700 tweets). This is an average improvement of 7.41% which is statistically significant. The best performing algorithms are Multilayer Perceptron and Random Forest. In order to be consistent to what was done previously, we used Multilayer Perceptron for the following experiments. With this classifier, the accuracy increased from 88.89% to 95.28%. This is a statistically significant improvement of 6.39%. Therefore, we can conclude that the more tweets are in the corpus, the better is the classifiers performance.

Figure 13 shows the improvement of each increment in the number of tweets per author in steps of 100. To simplify the visualisation, we only show the Multilayer Perceptron and the average of all the algorithms excluding Logistic Regression. We observe that the trend in both cases is clearly downward and tends to zero. On average, the highest decrease is from 300 to 400 tweets, whereas in case of Multilayer Perceptron it is from 600 to 700, and the slope is softer.

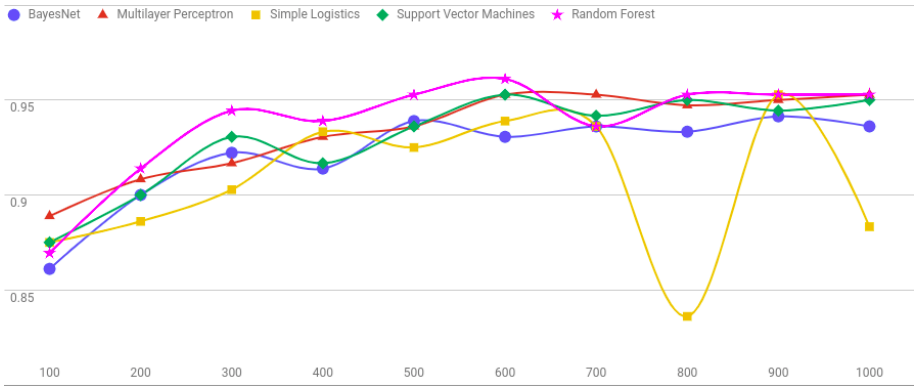

Fig. 12. Accuracy when the number of tweets increases (ARAP-Tweet corpus).

Figure 14 shows a decrease in accuracy when using less than 100 tweets. We observe that the trend is descending, slow at the beginning, and faster when the number of tweets decreases from 30. The average among classifiers decreases from 87.39% to 53% (i.e., by 34.39%), which is highly significant. In case of the Multilayer Perceptron, the decrease is higher, from 88.89% to 40.83%. However, this decrease in accuracy is not significant until the number of tweets is reduced to 40 for both the average classifier and Multilayer Perceptron.

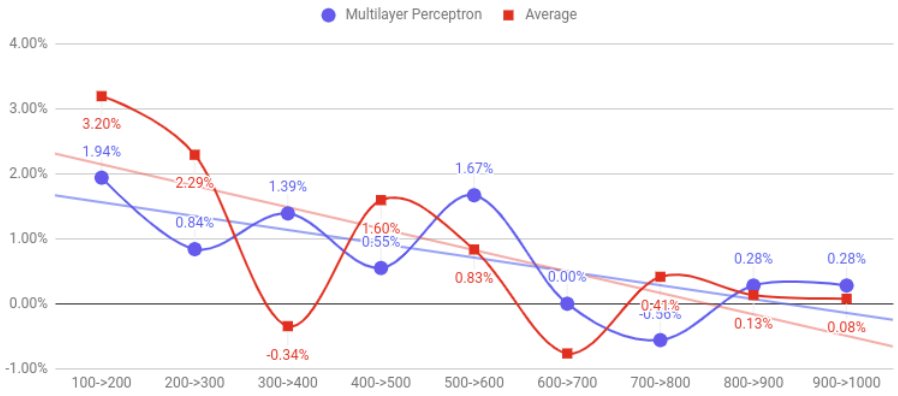

Fig. 13. Accuracy improvement for each increment in the number of tweets per user in steps of 100 (ARAP-Tweet corpus).

This analysis is important from the viewpoint of a real scenario because retrieving contents from Twitter and processing large amounts of tweets are both costly. Therefore, it is important to balance the quality with the cost, and to select the optimum number of tweets at which the accuracy improvement is not significant.

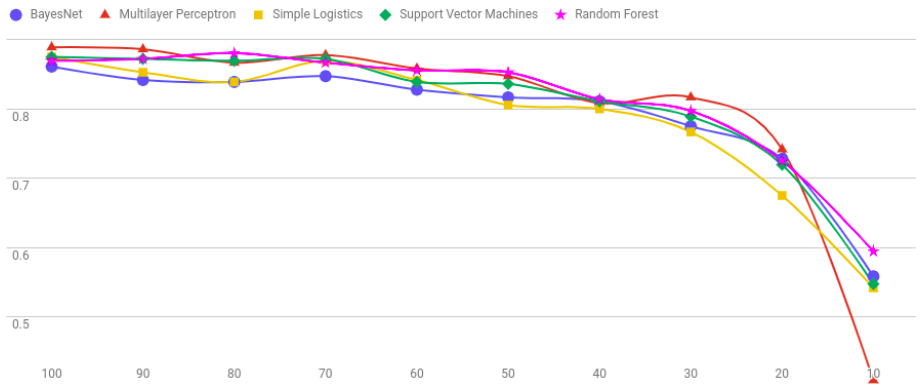

Fig. 14. Accuracy when the number of tweets per user decreases (ARAP-Tweet corpus).

6 Conclusions and future work

In this paper, we addressed the problem of fine-grained analysis of language varieties in the context of the authors demographics. We introduced the Low Dimensionality Statistical Embedding (LDSE) method that can be used to represent textual documents. We applied LDSE to the following two corpora: *(i)* the PAN-AP'17 corpus which covers four languages and includes the gender of their authors; *(ii)* the ARAP-Tweet corpus which covers 15 fine-grained Arabic varieties and includes the age and gender of their authors.

Our experiments with LDSE confirm its competitiveness with the state of the art. In fact, LDSE obtained an average accuracy of 92.08% over 91.84%, which was

obtained by the best performing team in the Author Profiling shared task at PAN 2017. We analysed the confusion among varieties, showing that usually the closer the regions are, the higher the confusion among their varieties is. We also analysed the variety identification error, considering the gender of the authors who wrote the tweets. We conclude that for the PAN-AP'17 corpus, the language variety of texts written by females is less difficult to be identified. We compared LDSE to the best performing teams at PAN and verified its competitiveness and stability. Based on that, we conclude that LDSE is very suitable for language variety identification.

We also analysed the performance of LDSE on the ARAP-Tweet corpus obtaining an average accuracy of 88.89% with Multilayer Perceptron. This result is more than 5% higher than the 83% obtained on the Arabic subset of the PAN-AP'17 corpus. We also analysed the confusion among varieties included in ARAP-Tweet obtaining similar results than previously. The closeness of regions increases the confusion among their language varieties. Moreover, we analysed the impact of the authors' age and gender on language variety identification. We conclude that in ARAP-Tweet it is less difficult to discriminate among varieties when the author is male, or when she/he belongs to the age classes *Between 25-34* or *Above 35*. We noticed strong differences compared to the results obtained at PAN with respect to the gender, which might be explained by the different methodologies used to build the two corpora. Finally, we analysed the impact of the corpus size on the classifiers performance, showing that the more tweets per user are in the corpus, the better the classifiers results are. Nevertheless, in a real scenario we should balance cost and performance.

As future work we will experiment in a grouped version of the ARAP-Tweet corpus. We will group the ARAP-Tweet corpus according to the regions defined by (Sadat *et al.* 2014) and we will apply LDSE and compare it with the results obtained with the PAN-AP'17 corpus. Furthermore, we will investigate the effect of cross-corpus evaluation. For that purpose, we will train with PAN-AP'17 and evaluate with ARAP-Tweet, and vice versa. This will allow us to know if these corpora can generalise well enough in order to be used in real application scenarios.

Acknowledgements

This publication was 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 authors.

References

Agić, Željko and Tiedemann, Jörg and Dobrovoljc, Kaja and Krek, Simon and Merkler, Danijela and Može, Sara and Nakov, Preslav and Osenova, Petya and Vertan, Cristina. Proceedings of the EMNLP 2014 Workshop on Language Technology for Closely Related Languages and Language Variants. Association for Computational Linguistics (2014)

Basile, Agenlo and Dwyer, Gareth and Medvedeva, Maria and Rawee, Josine and Haagsma, Hessel and Nissim, Malvina. Is there life beyond n-grams? a simple svm-based author profiling system. In 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 (2017)

Bogdanova, Dasha and Rosso, Paolo and Solorio, Thamar. Exploring high-level features for detecting cyberpedophilia. Computer speech & language, vol. 28 (1), pp. 108–120 Elsevier (2014)

Bowman, Kamiko O and Shenton, Leonard R. Method of moments. Encyclopedia of statistical sciences, vol. 5, pp. 467–473, John Wiley & Sons Canada (1985)

Castro, Dayvid and Souza, Ellen and de Oliveira, Adriano LI. Discriminating between Brazilian and European Portuguese national varieties on Twitter texts. In: 5th Brazilian Conference on Intelligent Systems (BRACIS), pp. 265–270 (2016)

Elfardy, Heba and Diab, Mona T. Sentence level dialect identification in arabic. Association for Computational Linguistics (ACL), pp. 456–461 (2013)

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

Gini, CW. Variability and mutability, contribution to the study of statistical distributions and relations. Studi cconomico-giuridici della r. Universita de cagliari (1912). Reviewed in: Light, rj, margolin, bh: An analysis of variance for categorical data." J. American Statistical Association 66 (1971): 534-544.

Grouin, Cyril and Forest, Dominic and Paroubek, Patrick and Zweigenbaum, Pierre. Présentation et résultats du défi fouille de texte DEFT2011 Quand un article de presse a t-il été écrit? À quel article scientifique correspond ce résumé? Actes du septième DÉfi Fouille de Textes, pp. 3 (2011)

Hagen, Matthias and Potthast, Martin and Stein, Benno. Overview of the Author Obfuscation Task at PAN 2018. CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org (2018)

Habash, Nizar. Introduction to Arabic natural language processing, vol. 3. Morgan & Claypool Publishers (2010)

Heitele, Dietger An epistemological view on fundamental stochastic ideas. Educational studies in Mathematics vol. 6, num. 2, pp. 187–205, Springer (1975)

Hernández-Fusilier, Donato and Montes-y-Gómez, Manuel and Rosso, Paolo and Cabrera-Guzmán, Rafael. Detecting positive and negative deceptive opinions using PU-learning. Information processing & management, vol. 51 (4), pp. 433–443 Elsevier (2015)

Huang, Chu-Ren and Lee, Lung-Hao. Contrastive approach towards text source classification based on top-bag-of-word similarity. PACLIC, pp. 404–410 (2008)

Inches, Giacomo and Crestani, Fabio. Overview of the International Sexual Predator Identification Competition at PAN-2012. CLEF Online working notes/labs/workshop, vol. 30 (2012)

Kandias, Miltiadis and Stavrou, Vasilis and Bozovic, Nick and Gritzalis, Dimitris. Proactive insider threat detection through social media: The YouTube case. In: Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society, pp. 261–266 (2013)

Kestemont, Mike and Tschuggnall, Michael and Stamatatos, Efstathios and Daelemans, Walter and Specht, Günther and Stein, Benno and Potthast, Martin. Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org (2018)

Lui, Marco and Cook, Paul. Classifying english documents by national dialect. In: Proceedings of the Australasian Language Technology Association Workshop, Citeseer pp. 5–15 (2013)

McNemar, Quinn. Note on the sampling error of the difference between correlated proportions or percentages. In: Psychometrika, 12(2), 153-157 (1947)

Maier, Wolfgang and Gómez-Rodrıguez, Carlos. Language variety identification in spanish tweets. LT4CloseLang (2014)

Malmasi, Shervin and Zampieri, Marcos and Ljubešić, Nikola and Nakov, Preslav and Ali, Ahmed and Tiedemann, Jörg. Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. In: Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), pp. 1–14 (2016)

Martinc, Matej and Skrjanec, Iza and Zupan, Katja and Pollak, Senja. Pan 2017: Author profiling - gender and language variety prediction. In 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 (2017)

Rangel, Francisco and Rosso, Paolo. On the impact of emotions on author profiling. Information processing & management, vol. 52 (1), pp. 73–92, Elsevier (2016a)

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

Rangel, Francisco and Rosso, Paolo and Potthast, Martin and Stein, Benno. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. Working Notes Papers of the CLEF 2017 Evaluation Labs, Editors: Linda Cappellato and Nicola Ferro and Lorraine Goeuriot and Thomas Mandl, pp. 1613–0073, CLEF and CEUR-WS.org (2017)

Rangel, and Rosso, Paolo and Montes-y-Gómez, Manuel and Potthast, Martin and Stein, Benno. Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org (2018)

Rosso, Paolo and Rangel, Francisco and Hernández-Farías, Irazu and Cagnina, Leticia and Zaghouani, Wajdi and Charfi, Anis. A survey on author profiling, deception, and irony detection for the Arabic language. Language and Linguistics Compass, vol. 12 (4), pp. e12275, Wiley Online Library (2018a)

Rosso, Paolo and Rangel Pardo, Francisco Manuel and Ghanem, Bilal and Charfi, Anis. ARAP: Arabic Author Profiling Project for Cyber-Security. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) (2018b)

Russell, Charles A and Miller, Bowman H. Profile of a terrorist. Studies in Conflict & Terrorism, vol. 1 (1), pp. 17–34, Taylor & Francis (1977)

Sadat, Fatiha and Kazemi, Farnazeh and Farzindar, Atefeh. Automatic identification of arabic language varieties and dialects in social media. Proceedings of SocialNLP, pp. 22 (2014)

Salton, Gerard and Buckley, Christopher: Term-weighting approaches in automatic text retrieval. Information processing & management, vol. 24(5), pp. 513–523 (1988)

Taylor, Robert W and Fritsch, Eric J and Liederbach, John. Digital crime and digital terrorism. Prentice Hall Press (2014)

Tellez, Eric S. and Miranda-Jiménez, Sabino and Graff, Mario and Moctezuma, Daniela. Gender and language variety identification with microtc. In 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 (2017)

Xu, Fan and Wang, Mingwen and Li, Maoxi. Sentence-level dialects identification in the Greater China region. International Journal on Natural Language Computing (IJNLC), 5(6) (2016)

Zaghouani, Wajdi and Charfi, Anis. ArapTweet: A Large MultiDialect Twitter Corpus for Gender, Age and Language Variety Identification. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan (2018a)

Zaghouani, Wajdi, and Charfi, Anis. Guidelines and Annotation Framework for Arabic Author Profiling. In Proceedings of the 3rd Workshop on Open-Source Arabic Corpora and Processing Tools, 11th International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan (2018b)

Zaidan, Omar F and Callison-Burch, Chris. Arabic dialect identification. Computational Linguistics, vol. 40 (1), pp. 171–202, MIT Press (2014)

Zampieri, Marcos and Gebre, Binyam Gebrekidan. Automatic identification of language varieties: the case of portuguese. In: The 11th conference on natural language processing (KONVENS), pp. 233–237 (2012)

Zampieri, Marcos and Tan, Liling and Ljubešić, Nikola and Tiedemann, Jörg. A report on the DSL shared task 2014. Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects, pp. 58–67 (2014)

Zampieri, Marcos and Tan, Liling and Ljubešić, Nikola and Tiedemann, Jörg and Nakov, Preslav. Overview of the dsl shared task 2015. Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects, pp. 1–9 (2015)

Zampieri, Marcos and Malmasi, Shervin and Ljubešić, Nikola and Nakov, Preslav and Ali, Ahmed and Tiedemann, Jörg and Scherrer, Yves and Aepli, Noëmi. Findings of the vardial evaluation campaign 2017. Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 1–15 (2017)