17th International Conference on Intelligent Text Processing and Computational Linguistics
April 3–9, 2016 • Konya, Turkey

A Low Dimensionality Representation for Language Variety Identification

Francisco Rangel[1,2], Marc Franco-Salvador[1], Paolo Rosso[1]
francisco.rangel@autoritas.es, mfranco@prhlt.upv.es, prosso@dsic.upv.es

[1]Universitat Politècnica de València, Spain
[2]Autoritas Consulting, Spain

Introduction

Language variety identification aims to detect linguistic variations in order to classify different varieties of the same language.

Language variety identification may be considered an **author profiling** task, besides a classification one, because the **cultural idiosyncrasies** may influence the way users use the language (e.g. different expressions, vocabulary…).

An example

The same sentence in different varieties of Spanish:

English	I was **goofing around** with my dog and I **lost** my **mobile**.
ES-Argentina	Estaba haciendo **boludeces** con mi perro y **extravié** el **celular**.
ES-Mexico	Estaba haciendo el **pendejo** con mi perro y **extravié** el **celular**.
ES-Spain	Estaba haciendo el **tonto** con mi perro y **perdí** el **móvil**.

Related Work

Tasks on language variety identification:

- Workshop on Language Technology for Closely Related Languages and Language Variants at EMNLP 2014
- VarDial Workshop - Applying NLP Tools to Similar Languages, Varieties and Dialects at COLING 2014
- T4VarDial - Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (DSL) shared task (Zampieri et al., 2014 and 2015) at RANLP 2015

Related Work

Authors	Varieties	Media	Features	Algorithm	Evaluation	Accuracy
Zampieri and Gebre (2012)	Portuguese	News	word and character n-grams	Probability distributions with log-likelihood	50-50 split	~90%
Sadat et al. (2014)	Arabic	Blogs Fora	character n-grams	Support Vector Machines	10-fold cross-validation	70-80%
Maier and Gómez-Rodríguez (2014)	Spanish	Twitter	character n-grams; LZW; syllable-based language models	Meta-learning	cross-validation	60-70%

Objective

To discriminate between different varieties of the same language, but with the following differences:

- We focus on different varieties of Spanish, although we tested our approach also with a different set of languages.
- Instead of n-gram based representations, we propose a low dimensionality representation which is helpful when dealing with big data in social media.
- We evaluate the proposed method with an independent test set generated from different authors in order to reduce possible overfitting.
- We make available our dataset to the research community. (https://github.com/autoritas/RD-Lab/tree/master/data/HispaBlogs)

Low dimensionality representation (LDR)

Step 1. Term-frequency - inverse document frequency (tf-idf) matrix:

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

- Each column is a vocabulary term *t*
- Each row is a document *d*
- w_{ij} is the tf-idf weight of the term *j* in the document i
- $\delta(d_i)$ represents the assigned class c to the document i

Step 2. Class-dependent term weighting:

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

Step 3. Class-dependent document representation:

$$d = \{F(c_1), F(c_2), \ldots, F(c_n)\} \sim \forall c \in C,$$

$$F(c_i) = \{avg, std, min, max, prob, prop\}$$

Low dimensionality representation (LDR)

avg	The average weight of a document is calculated as the sum of weights $W(t,c)$ of its terms divided by the total number of vocabulary terms of the document.
std	The standard deviation of the weight of a document is calculated as the root square of the sum of all the weights $W(t,c)$ minus the average.
min	The minimum weight of a document is the lowest term weight $W(t,c)$ found in the document.
max	The maximum weight of a document is the highest term weight $W(t,c)$ found in the document.
prob	The overall weight of a document is the sum of weights $W(t,c)$ of the terms of the document divided by the total number of terms of the document.
prop	The proportion between the number of vocabulary terms of the document and the total number of terms of the document.

Meaning of the measures

Alternative representations

- We use the common state-of-the-art representations based on n-grams. We iterated n from 1 to 10, and selected the 1000, 5000 and 10000 most frequent n-grams. The best results were obtained with:
  - character 4-grams; the 10,000 most frequent
  - word 1-gram (bag-of-words); the 10,000 most frequent
  - word 2-grams; the 10,000 highest tf-idf
- Two variations of the continuous Skip-gram model (Mikolov et al.):
  - Skip-grams
  - Sentence Vectors

Maximizing the average of the log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

Using the negative sampling estimator:

$$\log\sigma({v'_{w_O}}^T v_{w_I}) + \sum_{i=1}^{k}\mathbb{E}_{w_i}\sim P_n(w)\left[\log\sigma(-{v'_{w_i}}^T v_{w_I})\right]$$

Hispablogs dataset

Language Variety	# Blogs/authors		# Words		# Words per post	
	Training	Test	Training	Test	Training	Test
AR - Argentina	450	200	1,408,103	590,583	371 448	385 849
CL - Chile	450	200	1,081,478	298,386	313 465	225 597
ES - Spain	450	200	1,376,478	620,778	360 426	395 765
MX - Mexico	450	200	1,697,091	618,502	437 513	392 894
PE - Peru	450	200	1,602,195	373,262	410 466	257 627
TOTAL	2,250	1,000	7,164,935	2,501,511	380 466	334 764

Hispablogs dataset

- Completely independent authors between training and test sets
- Manually collected by social media experts of Autoritas

https://github.com/autoritas/RD-Lab/tree/master/data/HispaBlogs

Machine learning algorithms comparison

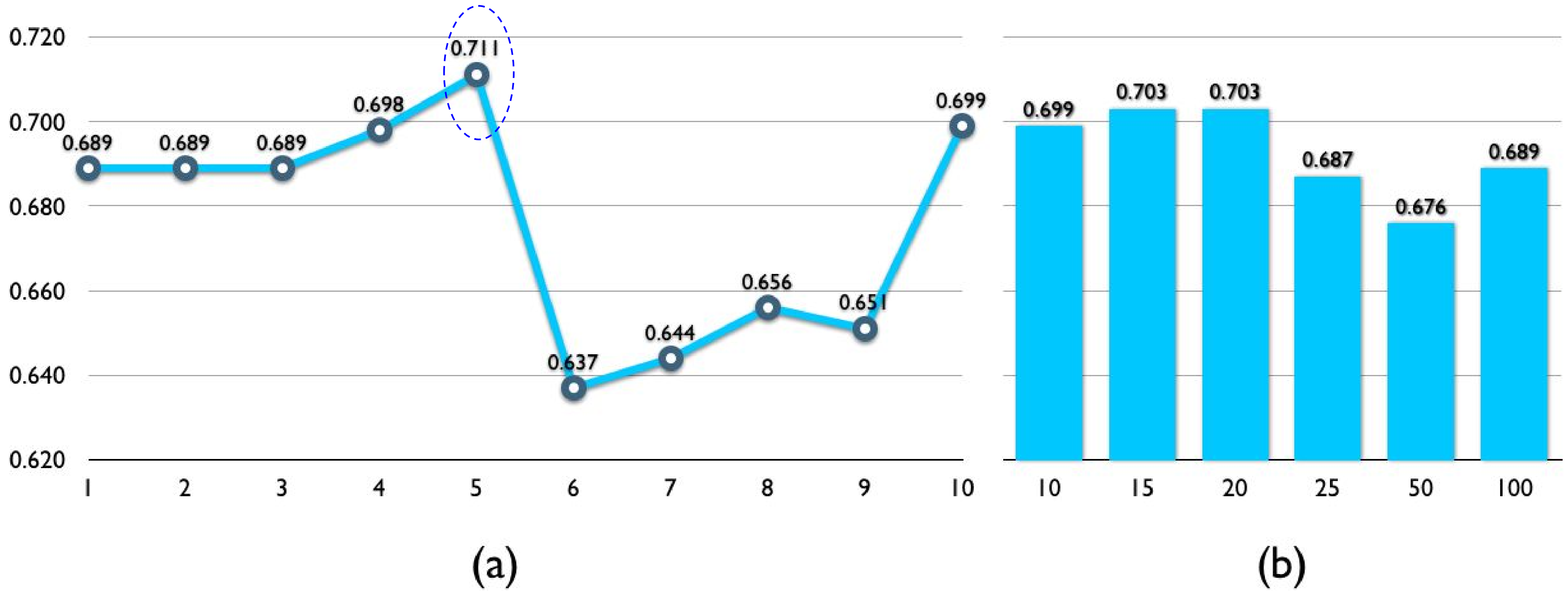
Algorithm	Accuracy	Algorithm	Accuracy	Algorithm	Accuracy
Multiclass Classifier	**71.1**	Rotation Forest	66.6	Multilayer Perceptron	62.5
SVM	69.3	Bagging	66.5	Simple Cart	61.9
LogitBoost	67.0	Random Forest	66.1	J48	59.3
Simple Logistic	66.8	Naive Bayes	64.1	BayesNet	52.2

Accuracy results with different machine learning algorithms

Significance of the results wrt. the two systems with the highest performance

$$\left\{ \begin{array}{l} \text{SVM } (z_{0.05}0,880 < 1,960) \\ \text{LogitBoost } (z_{0.05} = 1,983 > 1,960) \end{array} \right.$$

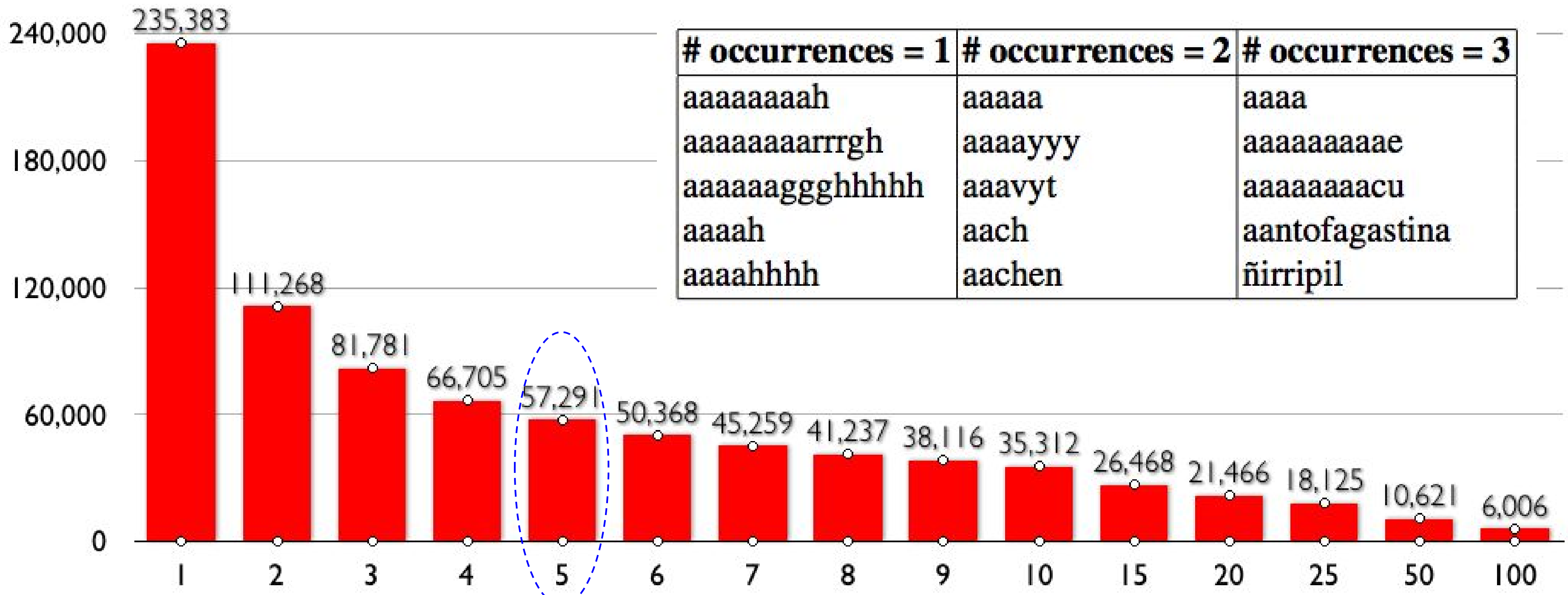
Preprocessing impact



Accuracy obtained after removing words with frequency equal or lower than n

(a) Continuous scale (b) Non-continuous scale

Preprocessing impact



# occurrences = 1	# occurrences = 2	# occurrences = 3
aaaaaaaah	aaaaa	aaaa
aaaaaaaarrrgh	aaaayyy	aaaaaaaaae
aaaaaaggghhhhh	aaavyt	aaaaaaaacu
aaaah	aach	aantofagastina
aaaahhhh	aachen	ñirripil

Number of words after removing those with frequency equal or lower than n, and some examples of very infrequent words.

Classification results

Representation	Accuracy
Skip-gram	0.722 *
LDR	**0.711**
SenVec	0.708 **
BOW	0.527
Char. 4-grams	0.515
tf-idf 2-grams	0.393
Random baseline	0.200

Accuracy results per representation

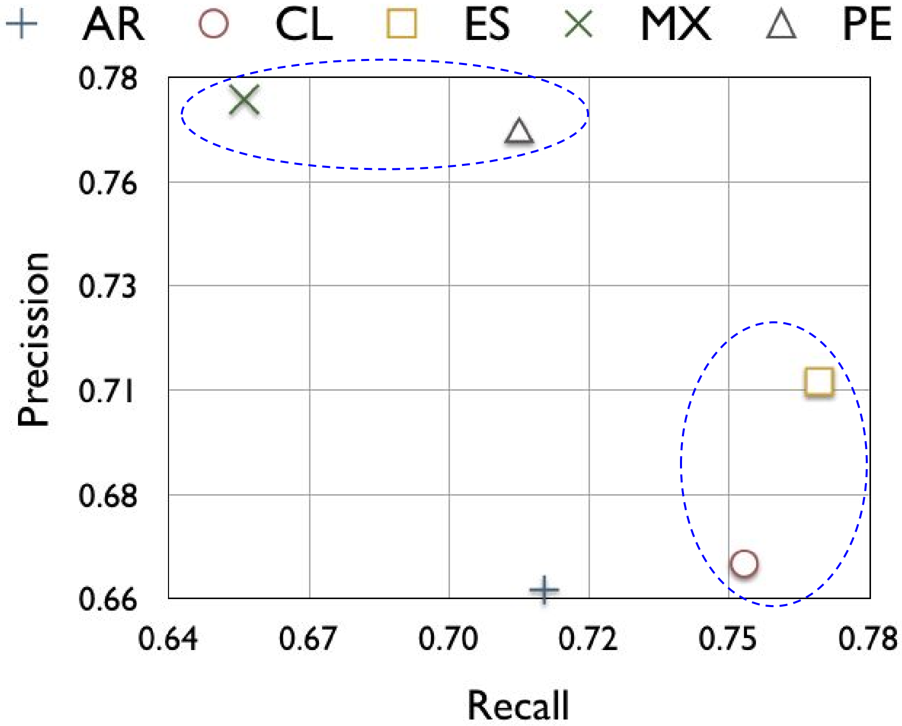
$^{*}z_{0.05} = 0,5457 < 1,960$

$^{**}z_{0.05} = 0,7095 < 1,960$

Error analysis

Variety	Clasified as AR	CL	ES	MX	PE
AR	143	16	22	8	11
CL	17	151	11	11	10
ES	20	13	154	7	6
MX	20	18	18	131	13
PE	16	28	12	12	132

Confusion matrix of the 5-class classification



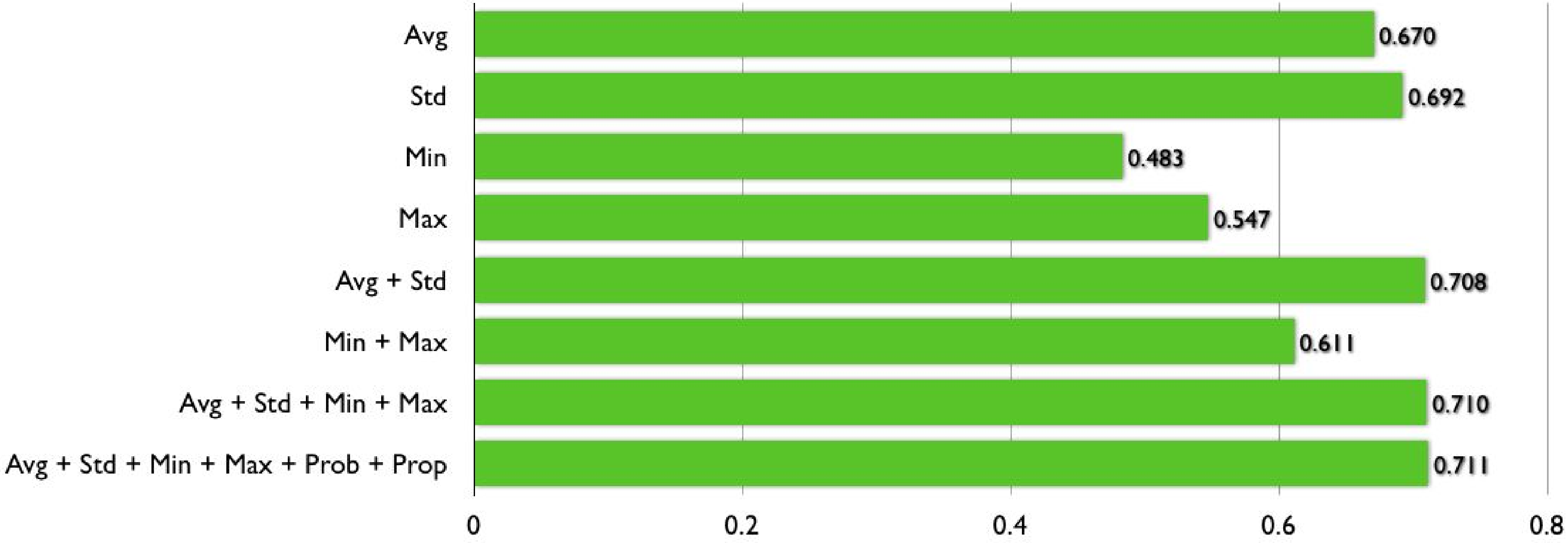
F1 values for identification as the corresponding language variety vs. others

Most discriminating features

Attribute	IG	Attribute	IG	Attribute	IG
PE-avg	0.680 ± 0.006	ES-std	0.497 ± 0.008	PE-prob	0.152 ± 0.005
AR-avg	0.675 ± 0.005	CL-max	0.496 ± 0.005	MX-prob	0.151 ± 0.005
MX-max	0.601 ± 0.005	CL-std	0.495 ± 0.007	ES-prob	0.130 ± 0.011
PE-max	0.600 ± 0.009	MX-std	0.493 ± 0.007	AR-prob	0.127 ± 0.006
ES-min	0.595 ± 0.033	CL-min	0.486 ± 0.013	AR-prop	0.116 ± 0.005
ES-avg	0.584 ± 0.004	AR-std	0.485 ± 0.005	MX-prop	0.113 ± 0.006
MX-avg	0.577 ± 0.008	PE-std	0.483 ± 0.012	PE-prop	0.112 ± 0.005
ES-max	0.564 ± 0.007	AR-min	0.463 ± 0.012	ES-prop	0.110 ± 0.007
AR-max	0.550 ± 0.007	CL-avg	0.455 ± 0.008	CL-prop	0.101 ± 0.005
MX-min	0.513 ± 0.027	PE-min	0.369 ± 0.019	CL-prob	0.087 ± 0.010

Features sorted by Information Gain

Most discriminating features



Accuracy obtained with different combinations of features

Cost analysis

Complexity of obtaining the features:

$$O(l \cdot n) + O(l \cdot m) = O(max(l \cdot n, l \cdot m)) = O(l \cdot n)$$

l: number of varieties
n: number of terms of the document
m: number of terms in the document that coincides with some term in the vocabulary
n $\geq$ m & l<<n

Number of features:

Representation	# Features
LDR	**30**
Skip-gram	300
SenVec	300
BOW	10,000
Char 4-grams	10,000
tf-idf 2-grams	10,000

Robustness

Results obtained with the development set of the **DSLCC** corpus from the **Discriminating between Similar Languages task** (2015)

Language	LDR	Skip-gram	SenVec
Bulgarian	99.9	100	100
Macedonian	99.9	100	100
Spain Spanish	**84.7**	82.1	**86.3**
Argentina Spanish	88.0	**90.3**	87.6
Portugal Portuguese	**87.4**	83.2	**90.0**
Brazilian Portuguese	**90.0**	**94.5**	87.6
Bosnian	**78.0**	80.3	74.4
Croatian	85.8	85.9	84.7
Serbian	**86.4**	75.1	91.2
Indonesian	99.4	99.3	99.4
Malay	99.2	99.2	**99.8**
Czech	99.8	99.9	99.8
Slovak	99.3	**100**	99.3
Other languages	99.9	99.8	99.8

NOTE: Significant results in bold

Conclusions and future work

LDR outperforms common state-of-the-art representations by **35%** increase in accuracy.

LDR obtains competitive results compared with two distributed representation-based approaches that employed the popular **continuous Skip-gram model**.

LDR remains competitive with different **languages** and **media** (DSLCC).

The **dimensionality reduction** is from thousands to only 6 features per language variety. This allows to deal with **big data** in **social media**.

We have applied LDR to **age and gender identification** and we plan to apply LDR to **personality recognition**.

Thank you very much!

Interested in digital text forensics (author profiling, authorship identification, author obfuscation)?
Do not hesitate and participate in the PAN laboratory!!

[***http://pan.webis.de/***](http://pan.webis.de/)

Francisco Rangel[1,2], Marc Franco-Salvador[1], Paolo Rosso[1]
francisco.rangel@autoritas.es, mfranco@prhlt.upv.es, prosso@dsic.upv.es
[1]Universitat Politècnica de València, Spain
[2]Autoritas Consulting, Spain