# Author Profiling
# Cross-genre evaluation

## PAN-AP-2016 CLEF 2016
## Évora, 5-8 September


Francisco Rangel
Autoritas Consulting

Paolo Rosso
Universitat Politècnica de Valencia

Ben Verhoeven & Walter Daelemans
University of Anwerp

Martin Potthast & Benno Stein
Bauhaus-Universität Weimar

# Introduction

**Author profiling** aims at identifying **personal traits** such as **age**, **gender**, personality traits, native language… from writings.

This is crucial for:

- Marketing
- Security
- Forensics

# Task goal

To investigate the effect of the **cross-genre** evaluation in the **age** and **gender** identification task.

Three languages:

English

Spanish

Dutch

# Corpus

**ENGLISH / SPANISH**

| Gender |
|---|
| Male vs. Female |
| **Age Groups** |
| 18-24; 25-34; 35-49; 50+ |

| | Training (Twitter) | | Early birds (Social Media) | | Test (Blogs) | |
|---|---|---|---|---|---|---|
| | English | Spanish | English | Spanish | English | Spanish |
| 18-24 | 26 | 16 | 70 | 16 | 10 | 4 |
| 25-34 | 136 | 64 | 92 | 20 | 24 | 12 |
| 35-49 | 182 | 126 | 102 | 16 | 32 | 26 |
| 50-64 | 78 | 38 | 80 | 8 | 10 | 10 |
| 65+ | 6 | 6 | 4 | 4 | 2 | 4 |
| Σ | 428 | 250 | 348 | 64 | 78 | 56 |

**DUTCH**

| Gender |
|---|
| Male vs. Female |

| Training (Twitter) | Early birds (Reviews) | Test (Reviews) |
|---|---|---|
| 384 | 50 | 500 |

# Evaluation measures

The **accuracy** is calculated per task and language.

Then, the averages per task are calculated:

$$\overline{gender} = \frac{gender\_en + gender\_es + gender\_nl}{3}$$

$$\overline{age} = \frac{age\_en + age\_es}{2}$$

$$\overline{joint} = \frac{joint\_en + joint\_es}{2}$$

Finally, the ranking is the global average:

$$ranking = \frac{\overline{gender} + \overline{age} + \overline{joint}}{3}$$

# Statistical significance

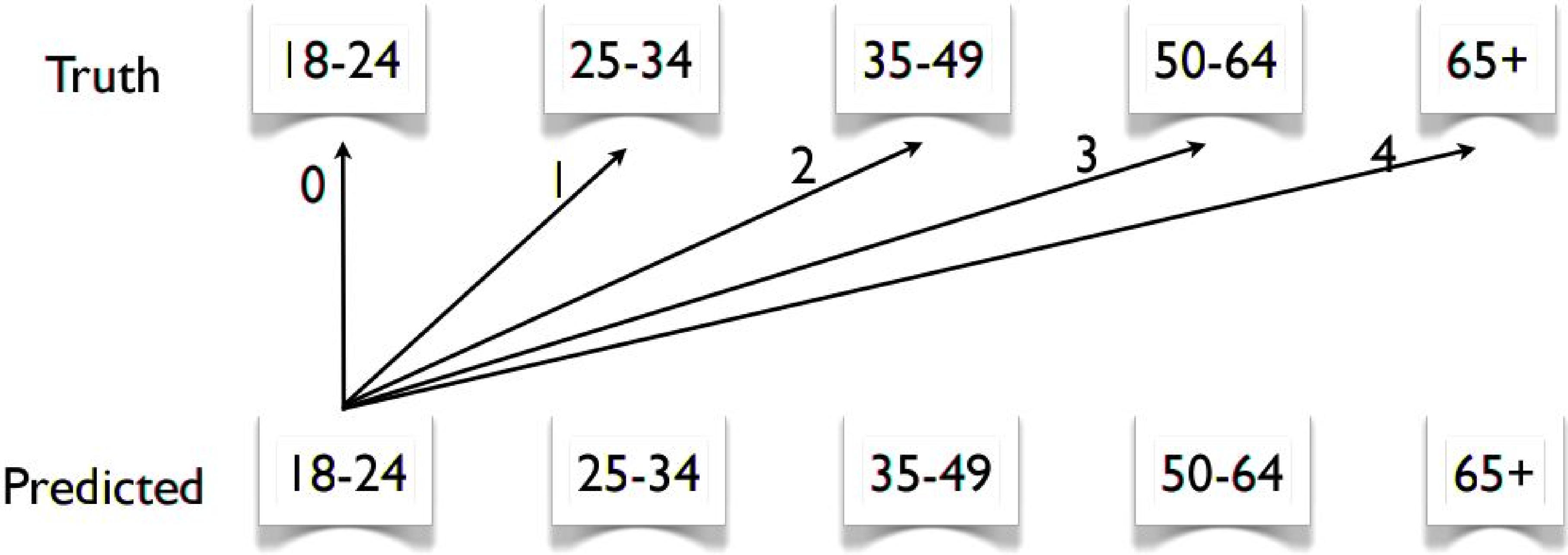
Approximate randomisation testing*

*Eric W. Noreen. Computer intensive methods for testing hypotheses: an introduction. Wiley, New York, 1989.

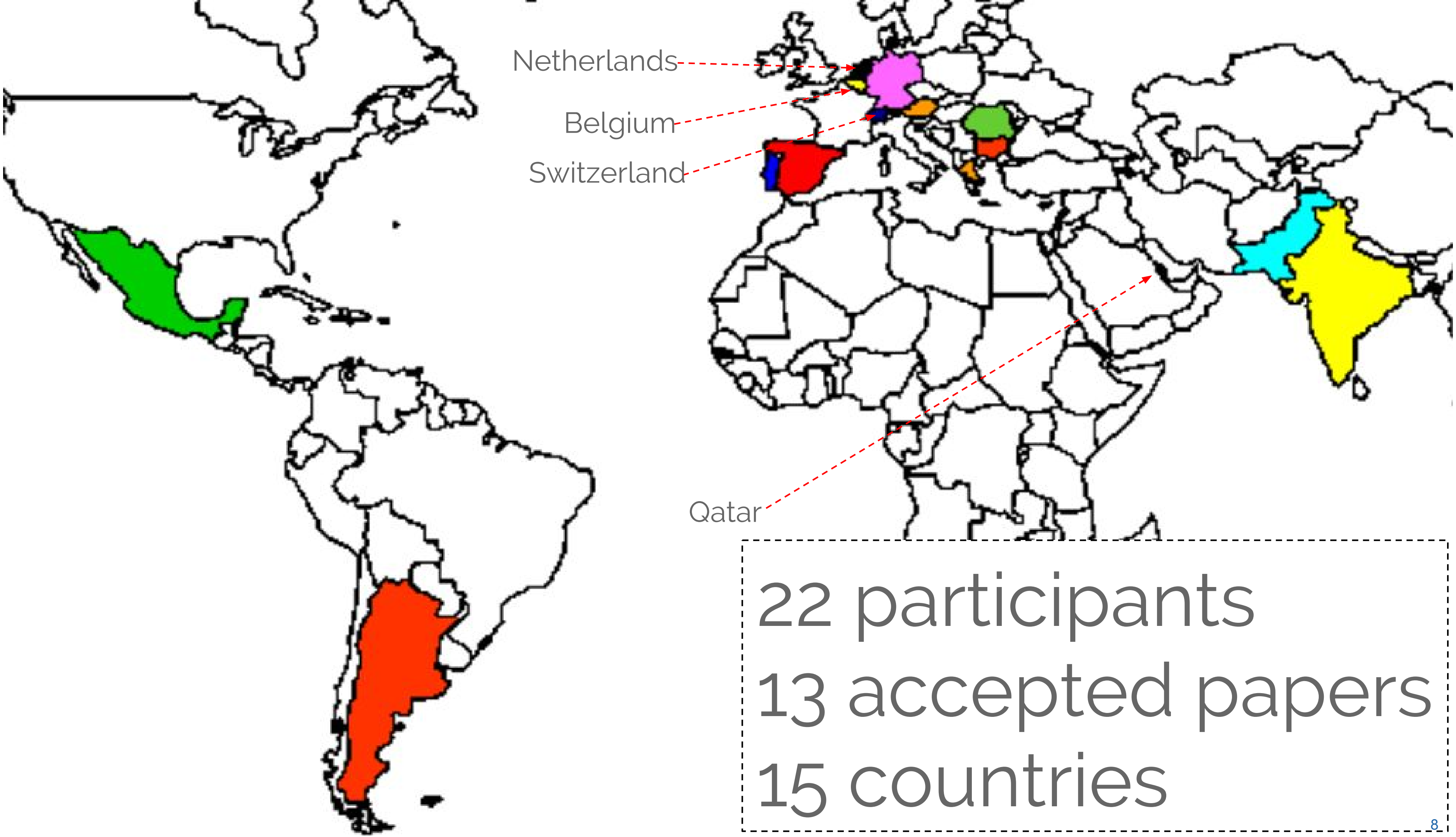
Pairwise comparison of accuracies of all systems

$p < 0.05$ -> the systems are significantly different

# Distances in age misidentification




- Missing predictions penalised with distance equal to 5
- Standard deviation of all the individual distances




22 participants
13 accepted papers
15 countries

# Approaches

What kind of ...

Preprocessing

Features

Methods

... did the teams perform?

# Approaches - Preprocessing

| | |
|---|---|
| HTML cleaning to obtain plain text | Devalkeener, Ashraf *et al.*, Bilan & Zhekova, Garciarena *et al.* |
| Lemmatization (no effect) | Bougiatiotis & Krithara |
| Stemming | Bakkar *et al.* |
| Punctuation signs | Bougiatiotis & Krithara, Gencheva *et al.*, Modaresi *et al.* |
| Stop words | Agrawal & Gonçalves, Bakkar *et al.* |
| Lowercase | Agrawal & Gonçalves, Bougiatiotis & Krithara |
| Digits removal | Bougiatiotis & Krithara, Markov *et al.* |
| Twitter specific components: hashtags, urls, mentions and RTs | Agrawal & Gonçalves, Bougiatiotis & Krithara, Markov *et al.*, Bilan & Zhekova, Kocher & Savoy, Gencheva *et al.* |
| Feature selection (no effect) | Ashraf *et al.*, Gencheva *et al.* |
| Transition point techniques | Markov *et al.* |

# Approaches - Features

| | |
|---|---|
| Stylistic features:<br>- Frequency of function words<br>- Words out of dictionary<br>- Slang<br>- Capital letters<br>- Unique words | Busger *et al.*, Ashraf *et al.*, Bougiatiotis & Krithara, Bilan & Zhekova, Gencheva *et al.*, Modaresi *et al.*, Pimas *et al.* |
| Specific sentences per gender<br>- My wife, my man, my girlfriend...<br>And per age<br>- "I'm" followed by a number | Gencheva *et al.* |
| Sentiment words | Gencheva *et al.*, Pimas *et al.* |
| N-gram models | Ashraf *et al.*, Bougiatiotis & Krithara, Modaresi *et al.*, Bilan & Zhekova, Gencheva *et al.*, Garciarena *et al.*, Markov *et al.* |
| Parts-of-speech | Bilan & Zhekova, Busger *et al.*, Gencheva *et al.*, Ashraf *et al.* |
| Collocations | Bilan & Zhekova |

# Approaches - Features

| | |
|---|---|
| LDA | Bilan & Zhekova |
| Different readability indexes | Gencheva *et al.* |
| Vocabulary richness | Ashraf *et al.* |
| Correctness | Pimas *et al.* |
| Verbosity | Dichiu & Rancea |
| Second order representation [22] | Busger *et al.*, Bougiatiotis & Krithara, Markov *et al.* |
| Bag-of-words | Devalkeener, Kocher & Savoy, Bakkar *et al.* |
| Tf-idf n-grams | Agrawal & Gonçalves, Dichiu & Rancea |
| Word2vec | Bayot & Gonçalves |

# Approaches - Methods

| | |
|---|---|
| Random Forest | Ashraf *et al.*, Pimas *et al.* |
| J48 | Ashraf *et al.* |
| LADTree | Ashraf *et al.* |
| Logistic regression | Modaresi *et al.*, Bilan & Zhekova |
| SVM | Bilan & Zhekova, Dichiu & Rancea, Bayot & Gonçalves, Markov *et al.*, Bougiatiotis & Krithara, Bakkar *et al.*, Busger *et al.* |
| SVM + bootstrap | Gencheva *et al.* |
| Stacking | Agrawal & Gonçalves |
| Class-RBM | Devalkeneer |
| Distance-based approaches | Kocher & Savoy, Garciarena *et al.* |

# Early birds evaluation in social media (EN/ES)

**English**

| Team | Joint | Gender | Age |
|---|---|---|---|
| Waser* | **0.2098** | 0.5230 | **0.3879** |
| Busger *et al.* | 0.1897 | **0.5575** | 0.3046 |
| Devalkeneer | 0.1839 | 0.5259 | 0.2931 |
| Dichiu & Rancea | 0.1753 | 0.5345 | 0.2989 |
| Agrawal & Gonçalves | 0.1724 | 0.5431 | 0.3103 |
| Bougiatiotis & Krithara | 0.1724 | 0.5345 | 0.3046 |
| Modaresi(a) | 0.1724 | 0.5057 | 0.3218 |
| Bilan *et al.* | 0.1667 | 0.5374 | 0.2902 |
| Gencheva *et al.* | 0.1638 | 0.5287 | 0.2902 |
| Garciarena *et al.* | 0.1609 | 0.5201 | 0.2816 |
| Kocher & Savoy | 0.1552 | 0.5144 | 0.2816 |
| Modaresi *et al.* | 0.1552 | 0.5029 | 0.3017 |
| Zahid | 0.1523 | 0.4885 | 0.3103 |
| Ashraf *et al.* | 0.1494 | 0.4971 | 0.2902 |
| Roman-Gomez | 0.1494 | 0.5144 | 0.2874 |
| Bakkar *et al.* | 0.1466 | 0.5029 | 0.2874 |
| *baseline* | 0.1207 | 0.5402 | 0.2126 |
| Pimas *et al.* | 0.0057 | 0.0201 | 0.0086 |

**Spanish**

| Team | Joint | Gender | Age |
|---|---|---|---|
| Bougiatiotis & Krithara | **0.2031** | 0.5781 | 0.3438 |
| Kocher & Savoy | **0.2031** | 0.5000 | 0.3125 |
| Modaresi *et al.* | **0.2031** | 0.6406 | 0.2813 |
| Busger *et al.* | 0.1875 | 0.5313 | 0.2813 |
| Devalkeneer | 0.1875 | 0.5625 | **0.3594** |
| Garciarena *et al.* | 0.1875 | 0.5625 | 0.2969 |
| Waser* | 0.1875 | **0.7031** | 0.2813 |
| Dichiu & Rancea | 0.1719 | 0.5469 | 0.2813 |
| Gencheva *et al.* | 0.1563 | 0.6250 | 0.2656 |
| Bilan *et al.* | 0.1406 | 0.5781 | 0.2969 |
| Modaresi(a) | 0.1406 | 0.6250 | 0.2969 |
| Zahid | 0.1406 | 0.5781 | 0.2969 |
| Agrawal & Gonçalves | 0.1094 | 0.4688 | 0.2500 |
| Roman-Gomez | 0.0938 | 0.5156 | 0.1563 |
| *baseline* | 0.0625 | 0.5313 | 0.1094 |

# Early birds evaluation in reviews (NL)

| Team | Gender |
|---|---|
| Roman-Gomez | **0.6200** |
| Waser* | 0.6000 |
| Gencheva *et al.* | 0.5600 |
| *baseline* | 0.5600 |
| Bayot & Gonçalves | 0.5400 |
| Bilan *et al.* | 0.5400 |

| Team | Gender |
|---|---|
| Dichiu & Rancea | 0.5400 |
| Garciarena *et al.* | 0.5400 |
| Zahid | 0.5400 |
| Kocher & Savoy | 0.5200 |
| Agrawal & Gonçalves | 0.5000 |
| Busger *et al.* | 0.5000 |

| Team | Gender |
|---|---|
| Devalkeneer | 0.5000 |
| Modaresi *et al.* | 0.5000 |
| Modaresi(a) | 0.5000 |
| Poongunran | 0.4800 |
| Bougiatiotis & Krithara | 0.4400 |

# Final evaluation in blogs (EN/ES)

| English | | | |
|---|---|---|---|
| Team | Joint | Gender | Age |
| Bougiatiotis & Krithara | **0.3974** | 0.6923 | 0.5513 |
| Busger *et al.* | 0.3846 | 0.6410 | **0.5897** |
| Modaresi *et al.* | 0.3846 | **0.7564** | 0.5128 |
| Bilan *et al.* | 0.3333 | 0.7436 | 0.4487 |
| Waser* | 0.3205 | 0.5897 | 0.4359 |
| Devalkeneer | 0.3205 | 0.6026 | 0.4487 |
| Modaresi(a) | 0.3205 | 0.6667 | 0.4487 |
| Markov *et al.* | 0.2949 | 0.6154 | 0.4487 |
| Roman-Gomez | 0.2821 | 0.6538 | 0.3974 |
| Dichiu & Rancea | 0.2692 | 0.6154 | 0.4103 |
| Gencheva *et al.* | 0.2564 | 0.6795 | 0.3718 |
| Kocher & Savoy | 0.2564 | 0.5769 | 0.4103 |
| Ashraf *et al.* | 0.2564 | 0.5769 | 0.3718 |
| Bayot & Gonçalves | 0.2179 | 0.6282 | 0.3590 |
| Deneva | 0.2051 | 0.5128 | 0.3718 |
| Bakkar *et al.* | 0.2051 | 0.5385 | 0.3718 |
| Agrawal & Gonçalves | 0.1923 | 0.5128 | 0.3846 |
| Zahid | 0.1923 | 0.5000 | 0.3846 |
| Aceituno | 0.1667 | 0.5000 | 0.3205 |
| Garciarena *et al.* | 0.1538 | 0.4615 | 0.3718 |
| Pimas *et al.* | 0.1410 | 0.5769 | 0.3205 |
| *baseline* | 0.0897 | 0.5641 | 0.1923 |

| Spanish | | | |
|---|---|---|---|
| Team | Joint | Gender | Age |
| Busger *et al.* | **0.4286** | 0.7143 | **0.5179** |
| Modaresi *et al.* | **0.4286** | 0.6964 | **0.5179** |
| Bilan *et al.* | 0.3750 | 0.6250 | 0.4643 |
| Markov *et al.* | 0.3750 | 0.6607 | 0.4464 |
| Dichiu & Rancea | 0.3214 | 0.6429 | 0.4643 |
| Bayot & Gonçalves | 0.3036 | 0.5893 | 0.4821 |
| Modaresi(a) | 0.3036 | 0.6964 | 0.4464 |
| Devalkeneer | 0.2857 | 0.5179 | 0.4821 |
| Agrawal & Gonçalves | 0.2857 | 0.5357 | 0.4821 |
| Deneva | 0.2679 | **0.7321** | 0.3214 |
| Waser* | 0.2679 | 0.5893 | 0.4107 |
| Bougiatiotis & Krithara | 0.2500 | 0.6786 | 0.3214 |
| Gencheva *et al.* | 0.2500 | 0.6250 | 0.3214 |
| Garciarena *et al.* | 0.2500 | 0.5000 | 0.4286 |
| Zahid | 0.2143 | 0.4821 | 0.4464 |
| Kocher & Savoy | 0.1964 | 0.5357 | 0.3393 |
| Roman-Gomez | 0.1250 | 0.5000 | 0.2500 |
| *baseline* | 0.1250 | 0.5000 | 0.1786 |
| Aceituno | 0.0893 | 0.4643 | 0.2143 |

# Final evaluation in reviews (NL)

| Team | Gender |
|---|---|
| Bayot & Gonçalves | **0.5680** |
| Roman-Gomez | 0.5620 |
| Bilan *et al.* | 0.5500 |
| Waser* | 0.5320 |
| *baseline* | 0.5300 |
| Dichiu & Rancea | 0.5260 |
| Garciarena *et al.* | 0.5260 |
| Poongunran | 0.5140 |
| Gencheva *et al.* | 0.5100 |
| Markov *et al.* | 0.5100 |
| Agrawal & Gonçalves | 0.5080 |
| Devalkeneer | 0.5060 |
| Aceituno | 0.5040 |
| Kocher & Savoy | 0.5040 |
| Modaresi *et al.* | 0.5040 |
| Busger *et al.* | 0.5000 |
| Modaresi(a) | 0.5000 |
| Deneva | 0.4980 |
| Bougiatiotis & Krithara | 0.4160 |

# Social media vs. blogs in English

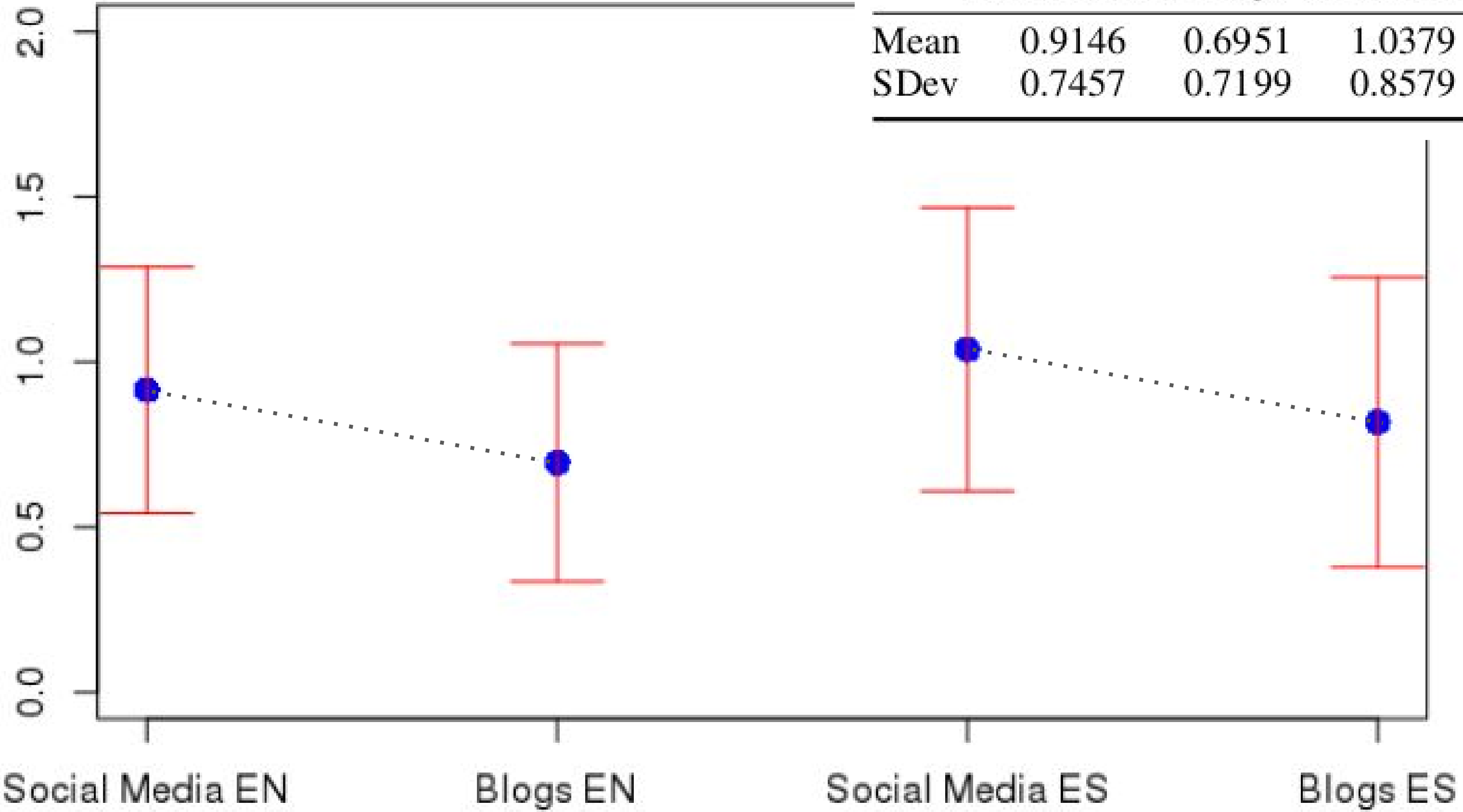
| | Joint | | Gender | | Age | |
|---|---|---|---|---|---|---|
| Team | Social Media | Blogs | Social Media | Blogs | Social Media | Blogs |
| Agrawal & Gonçalves | 0.1724 | 0.1923 | 0.5431 | *0.5128* | 0.3103 | 0.3846 |
| Ashraf *et al.* | 0.1494 | 0.2564 | 0.4971 | 0.5769 | 0.2902 | 0.3718 |
| Bakkar *et al.* | 0.1466 | 0.2051 | 0.5029 | 0.5385 | 0.2874 | 0.3718 |
| Bilan *et al.* | 0.1667 | 0.3333 | 0.5374 | 0.7436 | 0.2902 | 0.4487 |
| Bougiatiotis & Krithara | 0.1724 | **0.3974** | 0.5345 | 0.6923 | 0.3046 | 0.5513 |
| Busger *et al.* | 0.1897 | 0.3846 | **0.5575** | 0.6410 | 0.3046 | **0.5897** |
| Devalkeneer | 0.1839 | 0.3205 | 0.5259 | 0.6026 | 0.2931 | 0.4487 |
| Dichiu & Rancea | 0.1753 | 0.2692 | 0.5345 | 0.6154 | 0.2989 | 0.4103 |
| Garciarena *et al.* | 0.1609 | *0.1538* | 0.5201 | *0.4615* | 0.2816 | 0.3718 |
| Gencheva *et al.* | 0.1638 | 0.2564 | 0.5287 | 0.6795 | 0.2902 | 0.3718 |
| Kocher & Savoy | 0.1552 | 0.2564 | 0.5144 | 0.5769 | 0.2816 | 0.4103 |
| Modaresi(a) | 0.1724 | 0.3205 | 0.5057 | 0.6667 | 0.3218 | 0.4487 |
| Modaresi *et al.* | 0.1552 | 0.3846 | 0.5029 | **0.7564** | 0.3017 | 0.5128 |
| Pimas *et al.* | 0.0057 | 0.1410 | 0.0201 | 0.5769 | 0.0086 | 0.3205 |
| Roman-Gomez | 0.1494 | 0.2821 | 0.5144 | 0.6538 | 0.2874 | 0.3974 |
| Waser* | **0.2098** | 0.3205 | 0.5230 | 0.5897 | **0.3879** | 0.4359 |
| Zahid | 0.1523 | 0.1923 | 0.4885 | 0.5000 | 0.3103 | 0.3846 |
| Min | 0.0057 | 0.1410 | 0.0201 | 0.4615 | 0.0086 | 0.3205 |
| Q1 | 0.1523 | 0.2051 | 0.5029 | 0.5769 | 0.2874 | 0.3718 |
| Median | 0.1638 | 0.2692 | 0.5201 | 0.6026 | 0.2931 | 0.4103 |
| Mean | 0.1577 | 0.2745 | 0.4912 | 0.6109 | 0.2853 | 0.4253 |
| SDev | 0.0425 | 0.0794 | 0.1227 | 0.0827 | 0.0754 | 0.0704 |
| Q3 | 0.1724 | 0.3205 | 0.5345 | 0.6667 | 0.3046 | 0.4487 |
| Max | 0.2098 | 0.3974 | 0.5575 | 0.7564 | 0.3879 | 0.5897 |

# Social media vs. blogs in Spanish

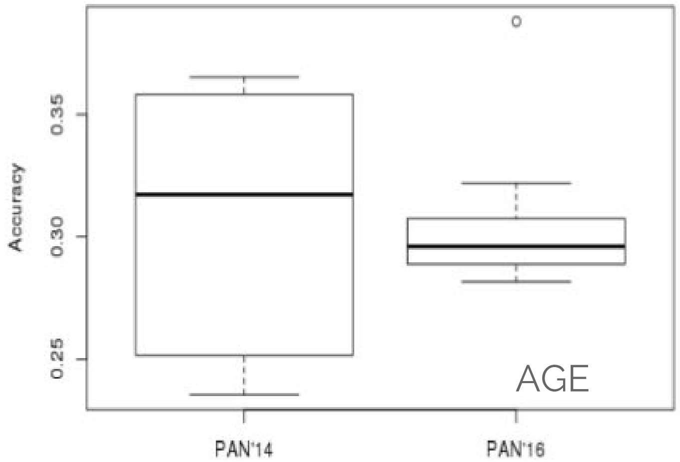
| | Joint | | Gender | | Age | |
|---|---|---|---|---|---|---|
| Team | Social Media | Blogs | Social Media | Blogs | Social Media | Blogs |
| Agrawal & Gonçalves | 0.1094 | 0.2857 | 0.4688 | 0.5357 | 0.2500 | 0.4821 |
| Bilan *et al.* | 0.1406 | 0.3750 | 0.5781 | 0.6250 | 0.2969 | 0.4643 |
| Bougiatiotis & Krithara | **0.2031** | 0.2500 | 0.5781 | 0.6786 | 0.3438 | *0.3214* |
| Busger *et al.* | 0.1875 | **0.4286** | 0.5313 | **0.7143** | 0.2813 | **0.5179** |
| Devalkeneer | 0.1875 | 0.2857 | 0.5625 | *0.5179* | **0.3594** | 0.4821 |
| Dichiu & Rancea | 0.1719 | 0.3214 | 0.5469 | 0.6429 | 0.2813 | 0.4643 |
| Garciarena *et al.* | 0.1875 | 0.2500 | 0.5625 | *0.5000* | 0.2969 | 0.4286 |
| Gencheva *et al.* | 0.1563 | 0.2500 | 0.6250 | 0.6250 | 0.2656 | 0.3214 |
| Kocher & Savoy | **0.2031** | *0.1964* | 0.5000 | 0.5357 | 0.3125 | 0.3393 |
| Modaresi(a) | 0.1406 | 0.3036 | 0.6250 | 0.6964 | 0.2969 | 0.4464 |
| Modaresi *et al.* | **0.2031** | **0.4286** | 0.6406 | 0.6964 | 0.2813 | **0.5179** |
| Roman-Gomez | 0.0938 | 0.1250 | 0.5156 | *0.5000* | 0.1563 | 0.2500 |
| Waser* | 0.1875 | 0.2679 | **0.7031** | *0.5893* | 0.2813 | 0.4107 |
| Zahid | 0.1406 | 0.2143 | 0.5781 | *0.4821* | 0.2969 | 0.4464 |
| Min | 0.0938 | 0.1250 | 0.4688 | 0.4821 | 0.1563 | 0.2500 |
| Q1 | 0.1406 | 0.2500 | 0.5352 | *0.5224* | 0.2813 | 0.3572 |
| Median | 0.1797 | 0.2768 | 0.5703 | 0.6072 | 0.2891 | 0.4464 |
| Mean | 0.1652 | 0.2844 | 0.5725 | 0.5957 | 0.2857 | 0.4209 |
| SDev | 0.0356 | 0.0848 | 0.0615 | 0.0831 | 0.0468 | 0.0819 |
| Q3 | 0.1875 | 0.3170 | 0.6133 | 0.6697 | 0.2969 | 0.4776 |
| Max | 0.2031 | 0.4286 | 0.7031 | 0.7143 | 0.3594 | 0.5179 |

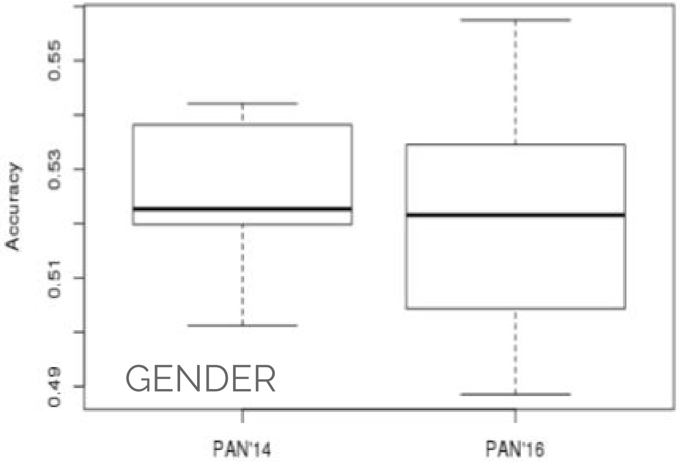
# Distances in age identification

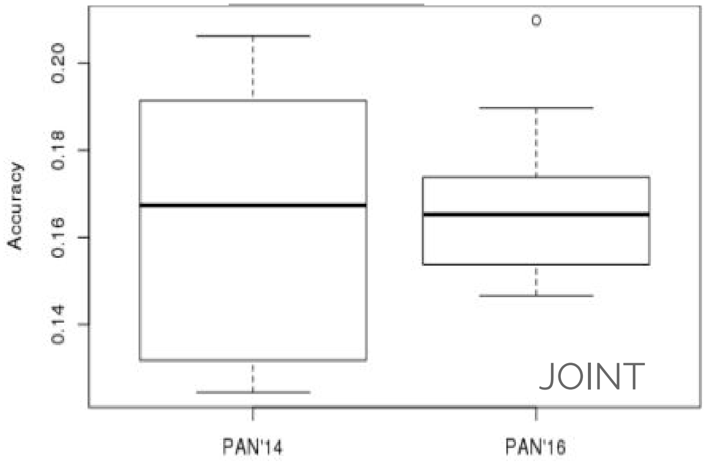
| | English | | Spanish | |
|---|---|---|---|---|
| | Social Media | Blogs | Social Media | Blogs |
| Mean | 0.9146 | 0.6951 | 1.0379 | 0.8176 |
| SDev | 0.7457 | 0.7199 | 0.8579 | 0.8775 |



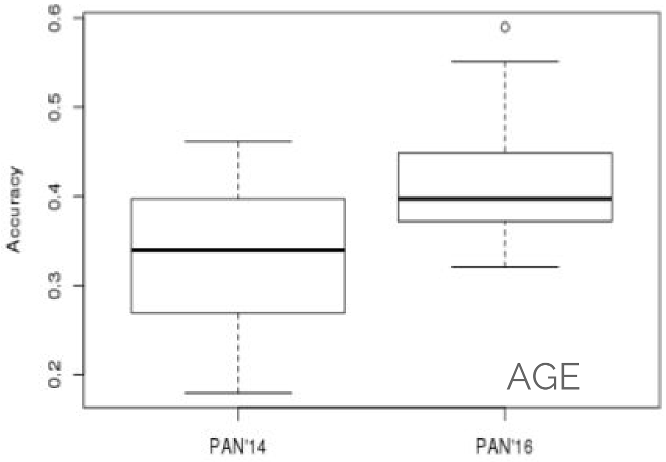
# 2014 vs. 2016 in social media (English)

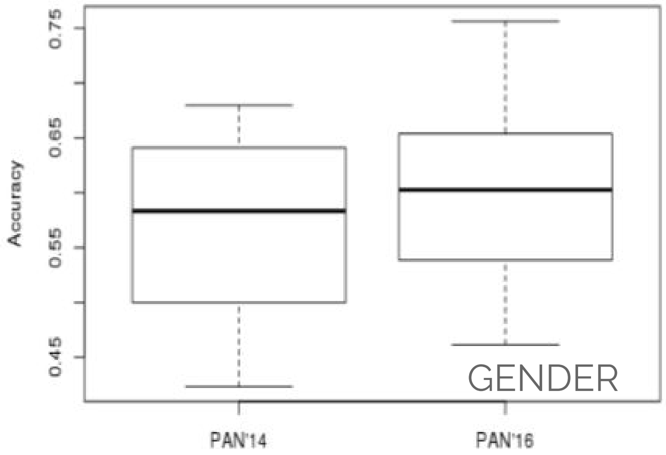


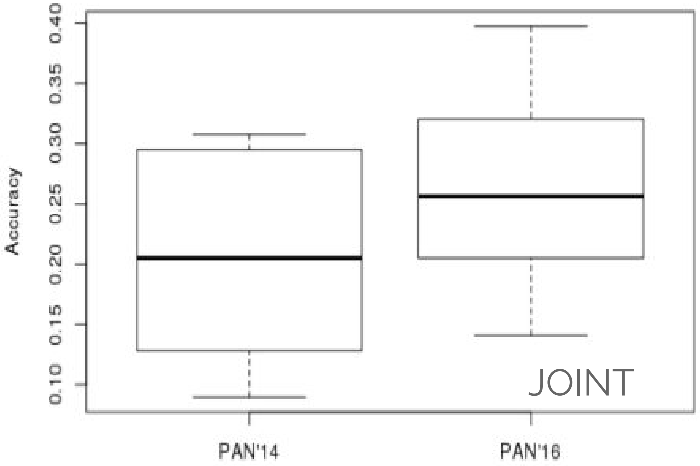



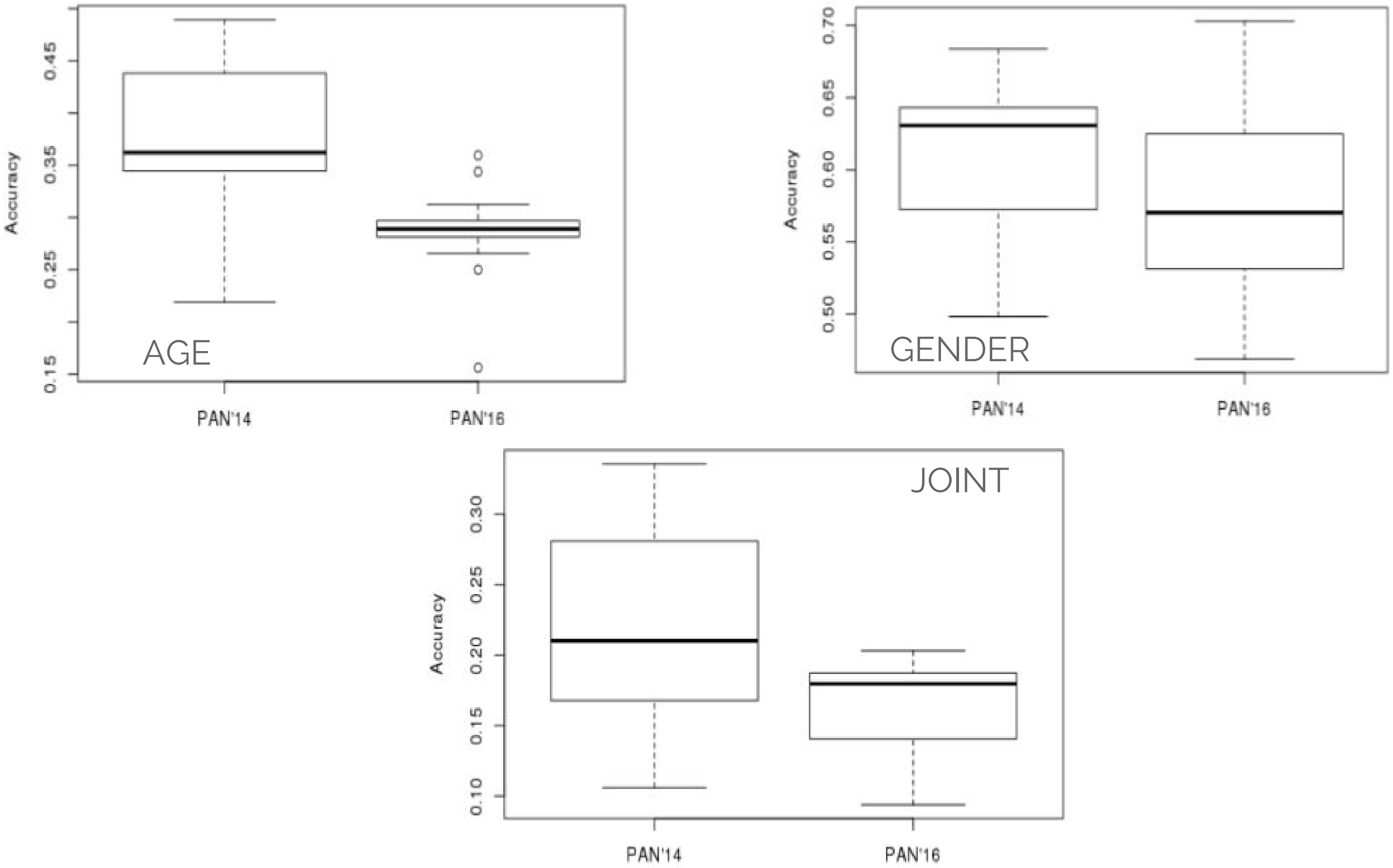
# 2014 vs. 2016 in blogs (English)



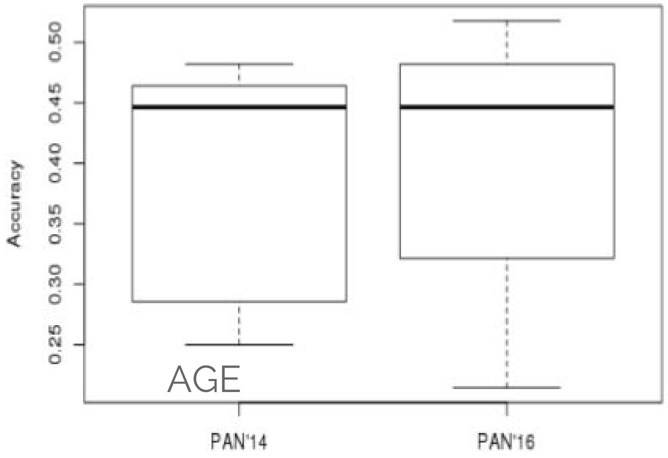



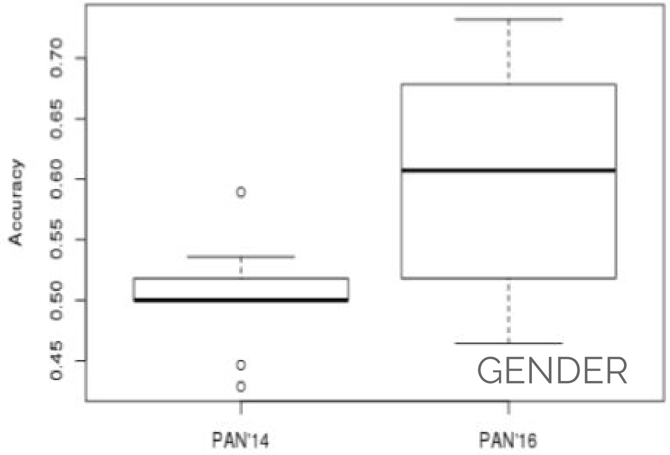



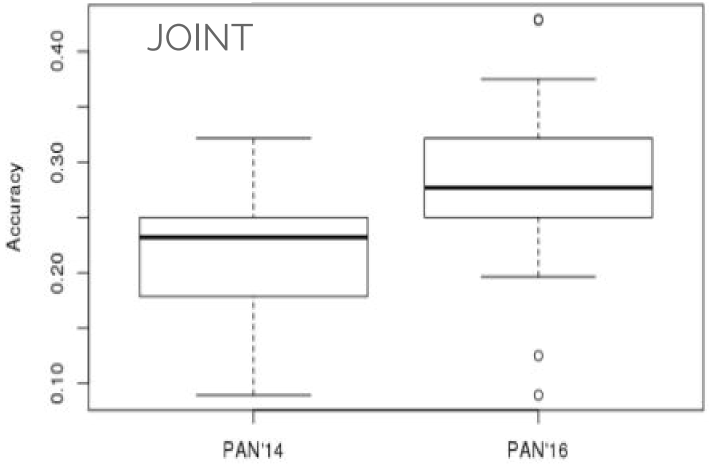
# 2014 vs. 2016 in social media (Spanish)



# 2014 vs. 2016 in blogs (Spanish)



# Final ranking

$$\overline{gender} = \frac{gender\_en + gender\_es + gender\_nl}{3}$$

$$\overline{age} = \frac{age\_en + age\_es}{2}$$

$$\overline{joint} = \frac{joint\_en + joint\_es}{2}$$

$$ranking = \frac{\overline{gender} + \overline{age} + \overline{joint}}{3}$$

| Ranking | Team | Global | English | Spanish | Dutch |
|---|---|---|---|---|---|
| 1 | Busger *et al.* | **0.5263** | 0.3846 | **0.4286** | 0.5000 |
| 2 | Modaresi *et al.* | 0.4934 | 0.3205 | **0.4286** | 0.5040 |
| 3 | Bilan *et al.* | 0.4834 | 0.3333 | 0.3750 | 0.5500 |
| 4 | Modaresi(a) | 0.4602 | 0.3205 | 0.3036 | 0.5000 |
| 5 | Markov *et al.* | 0.4593 | 0.2949 | 0.3750 | 0.5100 |
| 6 | Bougiatiotis & Krithara | 0.4519 | **0.3974** | 0.2500 | 0.4160 |
| 7 | Dichiu & Rancea | 0.4425 | 0.2692 | 0.3214 | 0.5260 |
| 8 | Devalkeneer | 0.4387 | 0.3205 | 0.2968 | 0.5060 |
| 9 | Waser* | 0.4293 | 0.3205 | 0.2679 | 0.5320 |
| 10 | Bayot & Gonçalves | 0.4255 | 0.2179 | 0.3036 | **0.5680** |
| 11 | Gencheva *et al.* | 0.4015 | 0.2564 | 0.2500 | 0.5100 |
| 12 | Agrawal & Gonçalves | 0.3971 | 0.1923 | 0.2857 | 0.5080 |
| 13 | Deneva | 0.3880 | 0.2051 | 0.2679 | 0.4980 |
| 14 | Kocher & Savoy | 0.3800 | 0.2564 | 0.1964 | 0.5040 |
| 15 | Roman-Gomez | 0.3664 | 0.2821 | 0.1250 | 0.5620 |
| 16 | Garciarena *et al.* | 0.3660 | 0.1538 | 0.2500 | 0.5260 |
| 17 | Zahid | 0.3154 | 0.1923 | 0.2143 | - |
| 18 | Aceituno | 0.2949 | 0.1667 | 0.0893 | 0.5040 |
| 19 | Poongunran | 0.1793 | - | - | 0.5140 |
| 20 | Ashraf *et al.* | 0.1688 | 0.2564 | - | - |
| 21 | Bakkar *et al.* | 0.1560 | 0.2051 | - | - |
| 22 | Pimas *et al.* | 0.1410 | 0.1410 | - | - |

# PAN-AP 2016 best results

| | Age and Gender | | |
|---|---|---|---|
| Language | *Joint* | Gender | Age |
| English | 0.3974 | 0.7564 | 0.5897 |
| Spanish | 0.4286 | 0.7321 | 0.5179 |
| Dutch | - | 0.5680 | - |

# Conclusions

- High combination of features: stylometric, n-grams, POS, collocations… First positions with:
    - Second order representation
    - Word2vec
- Early birds (social media in English and Spanish; reviews in Dutch):
    - Higher results for gender identification in Spanish than in English.
    - In Dutch and English most participants below baseline.
- Final evaluation (blogs in English and Spanish; reviews in Dutch):
    - Similar results for English and Spanish.
    - Most Dutch results below baseline.
- The effect of the cross-genre evaluation is higher in social media than in blogs:
    - Results in blogs are higher than in social media, except in case of gender identification in Spanish.
    - Distances in age identification are lower in blogs than in social media.
- Comparative results between 2014 and 2015 suggests:
    - There is no strong effect in the cross-genre evaluation in social media in English.
    - There is a strong impact in Spanish social media, specially in joint and age identification.
    - In blogs the effect is positive on age and joint identification in English and gender and joint in Spanish.
- Depending on the genre, the cross-genre may have a positive effect:
    - Learning from Twitter: spontaneous, without censorship, high number of tweets per user.
    - Evaluating on Blogs: difficult to obtain good labeled data.

# Task impact

| | PARTICIPANTS | COUNTRIES | CITATIONS |
|---|---|---|---|
| PAN-AP 2013 | 21 | 16 | 67 (+28) |
| PAN-AP 2014 | 10 | 8 | 41 (+25) |
| PAN-AP 2015 | 22 | 13 | 42 (+25) |
| PAN-AP 2016 | 22 | 15 | 5 |

# Industry at PAN (Author Profiling)

Organisation

autoritas

nuevas ideas, nuevas soluciones

Sponsors

meaning cloud

Adobe

Participants

KNOW Center

Bitdefender

# Next year?

On behalf of the author profiling task organisers:

Thank you very much for participating
and hope to see you next year!!