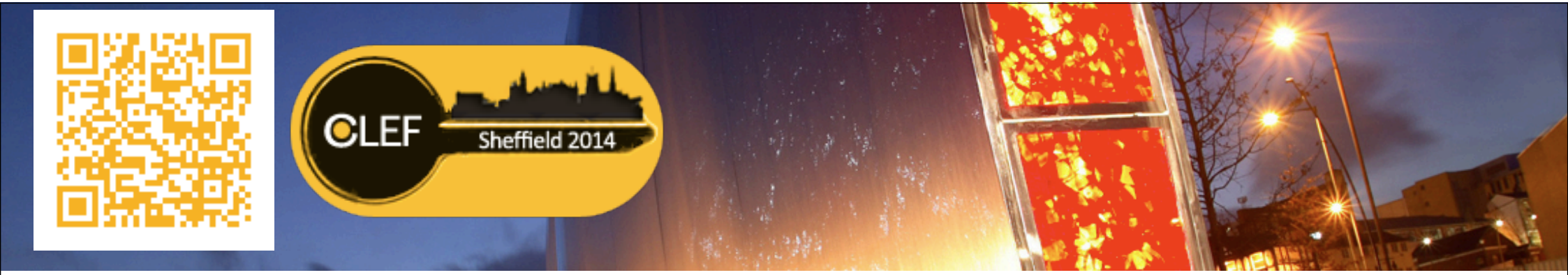


Author Profiling

PAN-AP-2014 - CLEF 2014

Sheffield, 15-18 September 2014

PAN

Francisco Rangel
Autoritas / Universitat Politècnica de València

Paolo Rosso
Universitat Politècnica de València

Irina Chugur
UNED

Martin Potthast, Martin Trenkmann, Benno Stein
Bauhaus-Universität Weimar

Ben Verhoeven, Walter Daelemans
University of Anwerp

What's Author Profiling?

Gender?

Personality traits?

Age?

Emotions?

Native language?

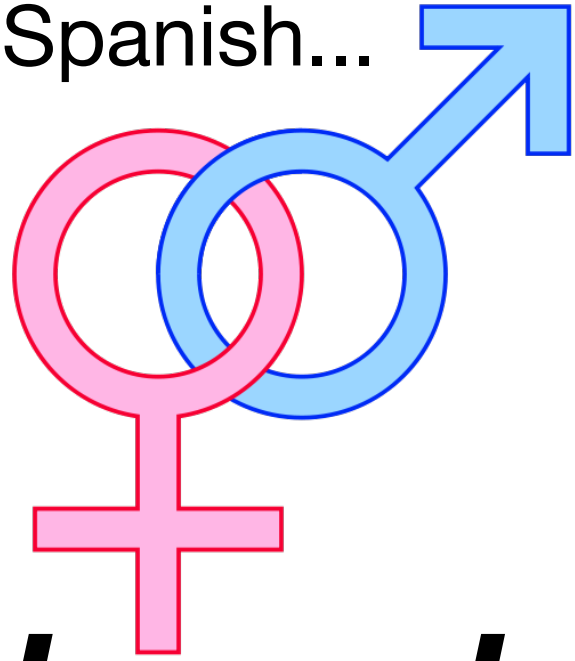
Author Profile... Who is who?

Why Author Profiling?

Forensics	Security	Marketing
Language as evidence	***Profile possible delinquents***	***Segmenting users***

Task Goal

- Given a collection of documents retrieved from different Social Media in English and Spanish...



To identify age and gender

Related Work on Author Profiling (age & gender)

AUTHOR	COLLECTION	FEATURES	RESULTS	OTHER CHARACTERISTICS
Argamon et al., 2002	British National Corpus	Part-of-speech	Gender: 80% accuracy	
Holmes & Meyerhoff, 2003	Formal texts	-	Age and gender	
Burger & Henderson, 2006	Blogs	Posts length, capital letters, punctuations. HTML features.	They only reported: "Low percentage errors"	Two age classes: [0,18[,[18,-]
Koppel et al., 2003	Blogs	Simple lexical and syntactic functions	Gender: 80% accuracy	Self-labeling
Schler et al., 2006	Blogs	Stylistic features + content words with the highest information gain	Gender: 80% accuracy Age: 75% accuracy	
Goswami et al., 2009	Blogs	Slang + sentence length	Gender: 89.18 accuracy Age: 80.32 accuracy	
Zhang & Zhang, 2010	Segments of blog	Words, punctuation, average words/sentence length, POS, word factor analysis	Gender: 72,10 accuracy	
Nguyen et al., 2011 y 2013	Blogs & Twitter	Unigrams, POS, LIWC	Correlation: 0.74 Mean absolute error: 4.1 - 6.8 years	Manual labeling Age as continuous variable
Peersman et al., 2011	Netlog	Unigrams, bigrams, trigrams and tetagrams	Gender+Age: 88.8 accuracy	Self-labeling, min 16 plus 16,18,25

News on PAN-AP 2014

News on Author Profiling

PAN-Replab Collaboration

Two complementary perspectives of **Author Proflining**

PAN virtual machines for **RepLab** participants TIRA platform @ Weimar

- All participants with the same **computing power**
- Increases participants **engagement**
- Improves **Sustainability, Replicability** and **Reproducibility**
- Allows **cross-year evaluations**

New Datasets

- PAN-AP13 -> Social Media
- TripAdvisor (EN)
- Blogs
- Twitter (with Replab)

Difficulty of collecting data

- Big Data?
- High variety of themes
- Real people vs. Robots (chatbots)
- Multilingual: English + Spanish + ...
- Difficulty to obtain (automatically) good label data
- Manual annotation?






Corpus

Social Media	Blogs	Twitter	Hotel reviews
▸ Subset of PAN-AP13 ▸ N. words > 100 ▸ Manual review	▸ Manually annotated (3 independent annotations) ▸ Personal blogs ▸ Up to 25 posts ▸ Rss content	▸ Manually annotated (3 independent annotations) ▸ Personal accounts ▸ Up to 1000 tweets ▸ Tweet Id. ▸ Replab collaboration	▸ TripAdvisor ▸ N. words > 10 ▸ Manual review
English Spanish			English
Balanced by gender			
Age groups: 18-24; 25-34; 35-49; 50-64; 65+			

Corpus – Social Media

LANG AGE GENDER			NUMBER OF AUTHORS		
			TRAINING	EARLY BIRDS	TEST
EN	18-24	MALE / FEMALE	1,550	140	680
	25-34		2,098	180	900
	35-49		2,246	200	980
	50-64		1,838	160	790
	65+		14	12	26
∑			***7,746***	***692***	***3,376***
ES	18-24	MALE / FEMALE	330	30	150
	25-34		426	36	180
	35-49		324	28	138
	50-64		160	14	70
	65+		32	14	28
∑			***1,272***	***122***	***566***

Corpus – Blogs

LANG	AGE	GENDER	NUMBER OF AUTHORS		
			TRAINING	EARLY BIRDS	TEST
EN	18-24	MALE / FEMALE	6	4	10
	25-34		60	6	24
	35-49		54	8	32
	50-64		23	4	10
	65+		4	2	2
∑			***147***	***24***	***78***
ES	18-24	MALE / FEMALE	4	2	4
	25-34		26	4	12
	35-49		42	4	26
	50-64		12	2	10
	65+		4	2	2
∑			***88***	***14***	***56***

Corpus – Twitter

LANG	AGE	GENDER	NUMBER OF AUTHORS		
			TRAINING	EARLY BIRDS	TEST
EN	18-24	MALE / FEMALE	20	2	12
	25-34		88	6	56
	35-49		130	16	58
	50-64		60	4	26
	65+		8	2	2
∑			***306***	***30***	***154***
ES	18-24	MALE / FEMALE	12	2	4
	25-34		42	4	26
	35-49		86	12	46
	50-64		32	6	12
	65+		6	2	2
∑			***178***	***26***	***90***

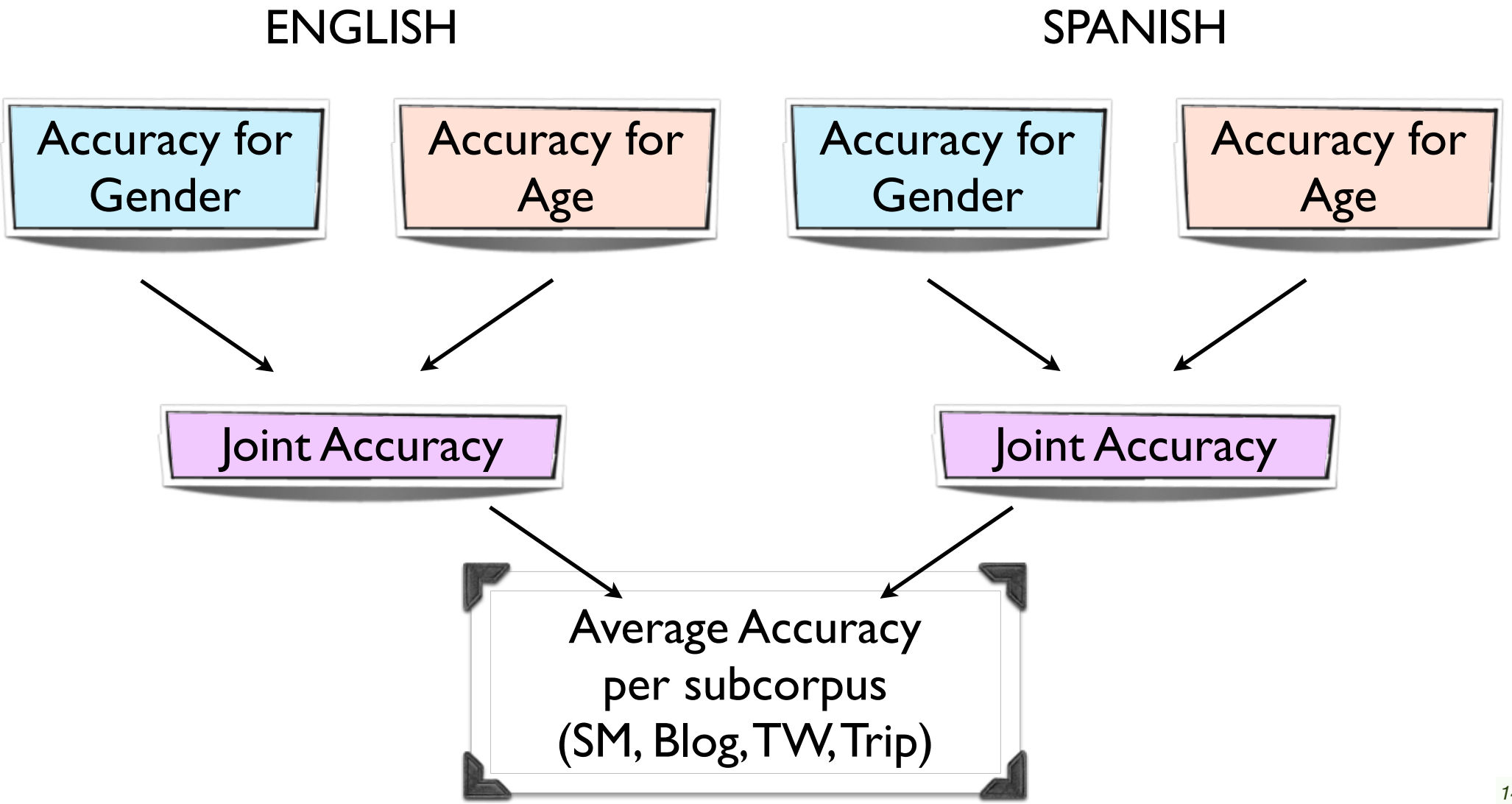
Corpus – Hotel reviews

LANG AGE GENDER			NUMBER OF AUTHORS	
			TRAINING	TEST
EN	18-24	MALE / FEMALE	180	74
	25-34		500	200
	35-49		500	200
	50-64		500	200
	65+		400	147
∑			***2,080***	***821***

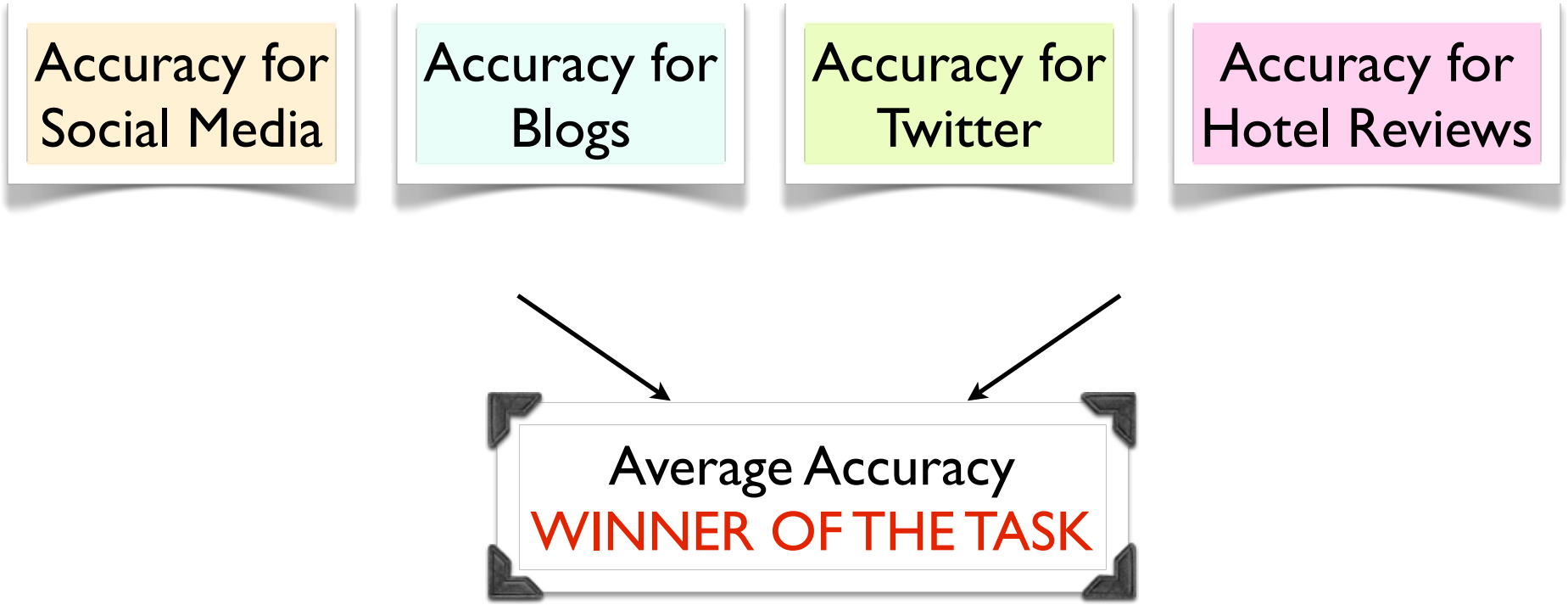
Corpus (test)

GENDER / AGE		SOCIAL MEDIA		BLOGS		TWITTER		REVIEWS
		EN	ES	EN	ES	EN	ES	EN
FEMALE	18-24	340	75	5	2	6	2	74
	25-34	450	90	12	6	28	13	200
	35-49	490	69	16	13	29	23	200
	50-64	395	35	5	5	13	6	200
	65+	13	14	1	1	1	1	147
MALE	18-24	340	75	5	2	6	2	86
	25-34	450	90	12	6	28	13	250
	35-49	490	69	16	13	29	23	302
	50-64	395	35	5	5	13	6	268
	65+	13	14	1	1	1	1	178
∑		***3376***	***566***	***78***	***56***	***154***	***90***	***1905***

Identification accuracies



Participants' ranking



BASELINE: The 1000 most frequent character trigrams with SVM

Statistical significance

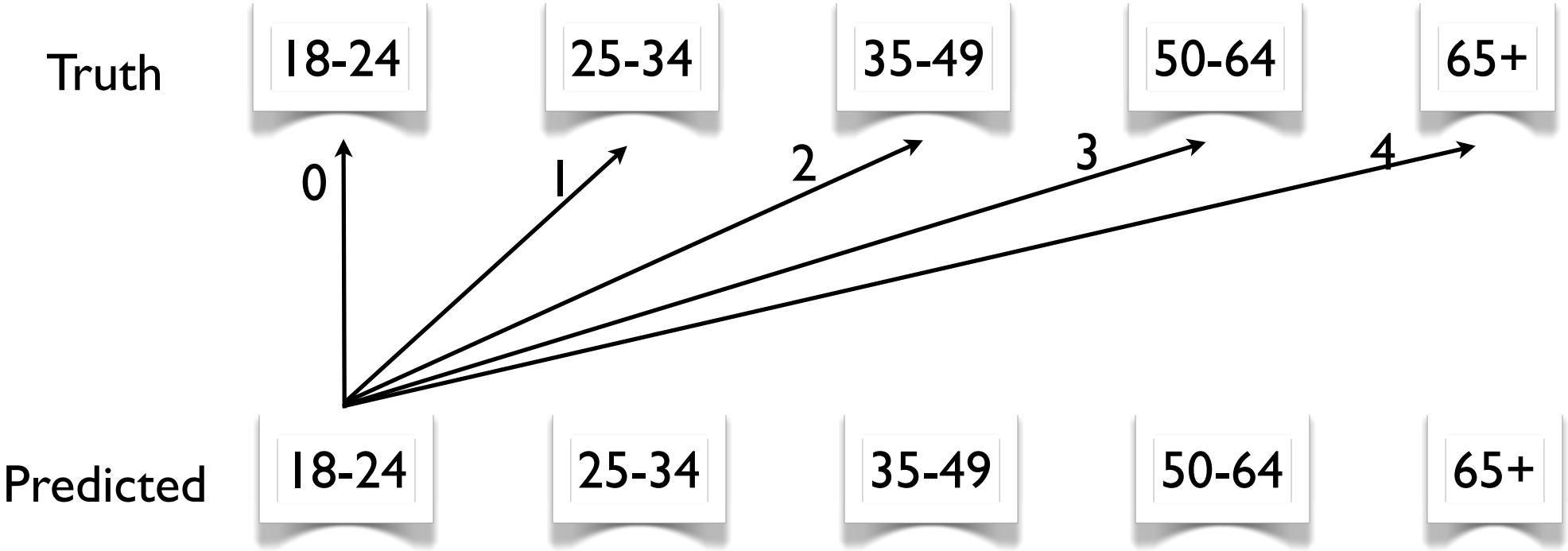
Approximate randomisation testing*

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

Pairwise comparison of accuracies of all systems

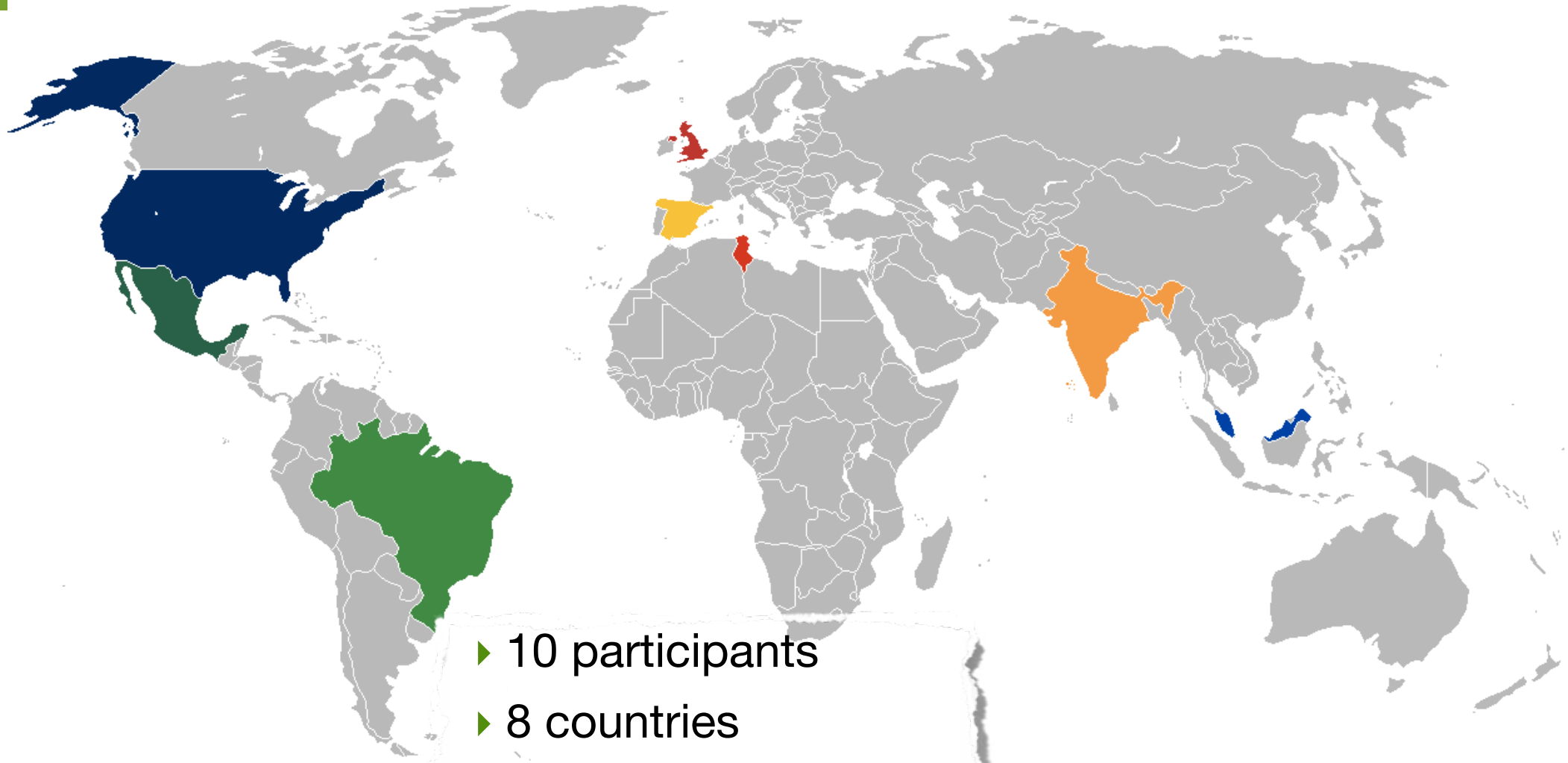
$p < 0.05$ -> the systems are significantly different

Distances in age misidentification



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

Participants



- 10 participants
- 8 countries
- 8 papers

Approaches

- What kind of ...

Preprocessing

Features

Methods

... did the teams perform?

Approaches

Preprocessing

HTML Cleaning to obtain plain text	5 teams: [shrestha][marquardt][baker] [ashok][weren]
Deletion of URLs, hashtags and user mentions in Twitter	1 team: [ashok]
Case conversion, invalid characters, multiple white spaces...	2 team: [baker][weren]
Tokenisation	2 teams: [villenaroman][weren]
Subset selection	1 team: [weren]
Discrimination between human-like posts and spam-like posts (chatbots)	1 team: [marquardt]

Approaches

Features

Stylistic features: frequencies of punctuation marks, size of sentences, words that appear once and twice, use of deflections, number of characters, words and sentences...	7 teams: [mechti][marquardt][ashok] [baker][weren][shrestha][liau]
Number of posts per user	1 team: [marquardt]
Correctness, cleanliness, diversity of texts	1 team: [weren]
HTML tags such as img, href, br	2 teams: [weren][marquardt]

Approaches

Features

Readability measures: Automated readability index, Coleman-Liau index, Rix Readability Index, Gunning Fog Index, Flesch-Kinkaid Index...	5 teams: [mechti][marquardt][ashok] [baker][weren]
Lexical Analysis: PoS, proper nouns, character flooding...	2 teams: [mechti][ashok]
Emoticons	3 teams: [shrestha][marquardt][liau]

Approaches

Features

Content features: n-grams, bag-of-words	3 teams: [villenaroman][shrestha][liau]
Topic words: money, home, smartphone...	1 team: [mechti]
MRC, LIWC: familiarity, concreteness, imagery, motion, emotion, religion...	1 team: [marquardt]
Dictionaries per subcorpus and class, lexical errors, foreign words, specific phrases: my husband, my wife...	4 teams: [baker][marquardt][ashok][liau]

Approaches

Features

Sentiment	1 team: [marquardt]
Text to be identified is used as a query for a search engine: cosine similarity, Okapi BM25	1 team: [weren]
Second order representation based on relationships among terms, documents, profiles and subprofiles	1 team: [pastor]

Approaches

Methods

Logistic Regression	1 team: [shrestha][liau][weren]
Logic Boost, Rotation Forest, Multi-Class Classifier, Multilayer Perceptron, Simple Logistic	1 team: [weren]
Multinomial Naïve Bayes	1 team: [villenaroman]
libLINEAR	1 team: [lopezmonroy]
Random Forest	1 team: [ashok]
Support Vector Machines	1 team: [marquardt]
Decision Tables	1 team: [mecthi]
Own Frequency-based Prediction Function	1 team: [baker]

Early birds (best) results

	ENGLISH			SPANISH		
CORPUS	JOINT	GENDER	AGE	JOINT	GENDER	AGE
SOCIAL MEDIA	liau (0.2153)	liau (0.5390)	liau (0.3728)	shrestha (0.3033)	liau (0.7295)	liau (0.4262)
BLOG	lopezmonroy (0.2083)	lopezmonroy (0.6250)	4 teams (0.2500)	lopezmonroy (0.3571)	marquardt (0.6429)	2 teams (0.4286)
TWITTER	lopezmonroy (0.5333)	lopezmonroy (0.7667)	lopezmonroy (0.6333)	shrestha (0.6154)	shrestha (0.8846)	shrestha (0.6923)
HOTEL REVIEWS	liau (0.2622)	liau (0.7317)	lopezmonroy (0.3720)	-		

- 7 teams participated

Final (best) results

	ENGLISH			SPANISH		
CORPUS	JOINT	GENDER	AGE	JOINT	GENDER	AGE
SOCIAL MEDIA	shrestha (0.2062)	villenaroman (0.5421)	shrestha (0.3652)	liau (0.3357)	liau (0.6837)	liau (0.4894)
BLOG	2 teams (0.3077)	lopezmonroy (0.6795)	weren (0.4615)	lopezmonroy (0.3214)	lopezmonroy (0.5893)	2 teams (0.4821)
TWITTER	lopezmonroy (0.3571)	liau (0.7338)	liau (0.5065)	shrestha (0.4333)	shrestha (0.6556)	shrestha (0.6111)
HOTEL REVIEWS	liau (0.2564)	liau (0.7259)	liau (0.3502)	-		

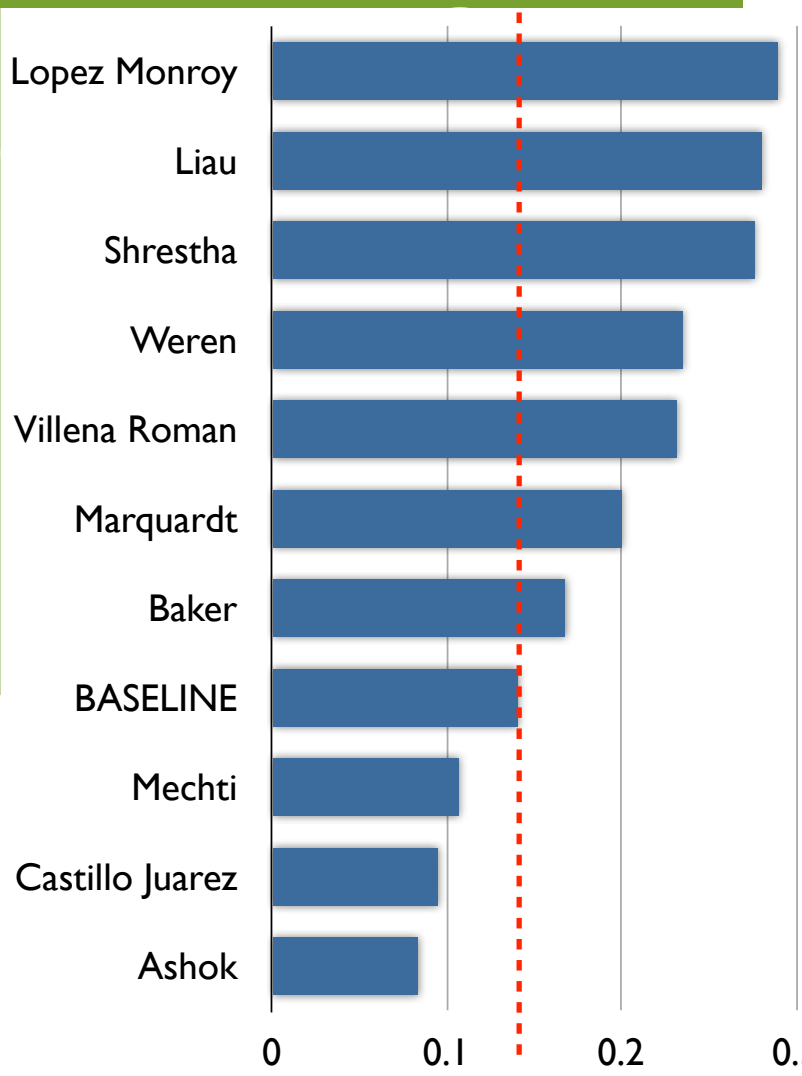
- 10 teams participated

Final (best) results

	ENGLISH			SPANISH		
CORPUS	JOINT	GENDER	AGE	JOINT	GENDER	AGE
SOCIAL MEDIA	shrestha (0.2062)	villenaroman (0.5421)	shrestha (0.3652)	liau (0.3357)	liau (0.6837)	liau (0.4894)
BLOG	2 teams (0.3077)	lopezmonroy (0.6795)	weren (0.4615)	lopezmonroy (0.3214)	lopezmonroy (0.5893)	2 teams (0.4821)
TWITTER	lopezmonroy (0.3571)	liau (0.7338)	liau (0.5065)	shrestha (0.4333)	shrestha (0.6556)	shrestha (0.6111)
HOTEL REVIEWS	liau (0.2564)	liau (0.7259)	liau (0.3502)	-		

- High performance of the content features: n-grams, BoW

Average results



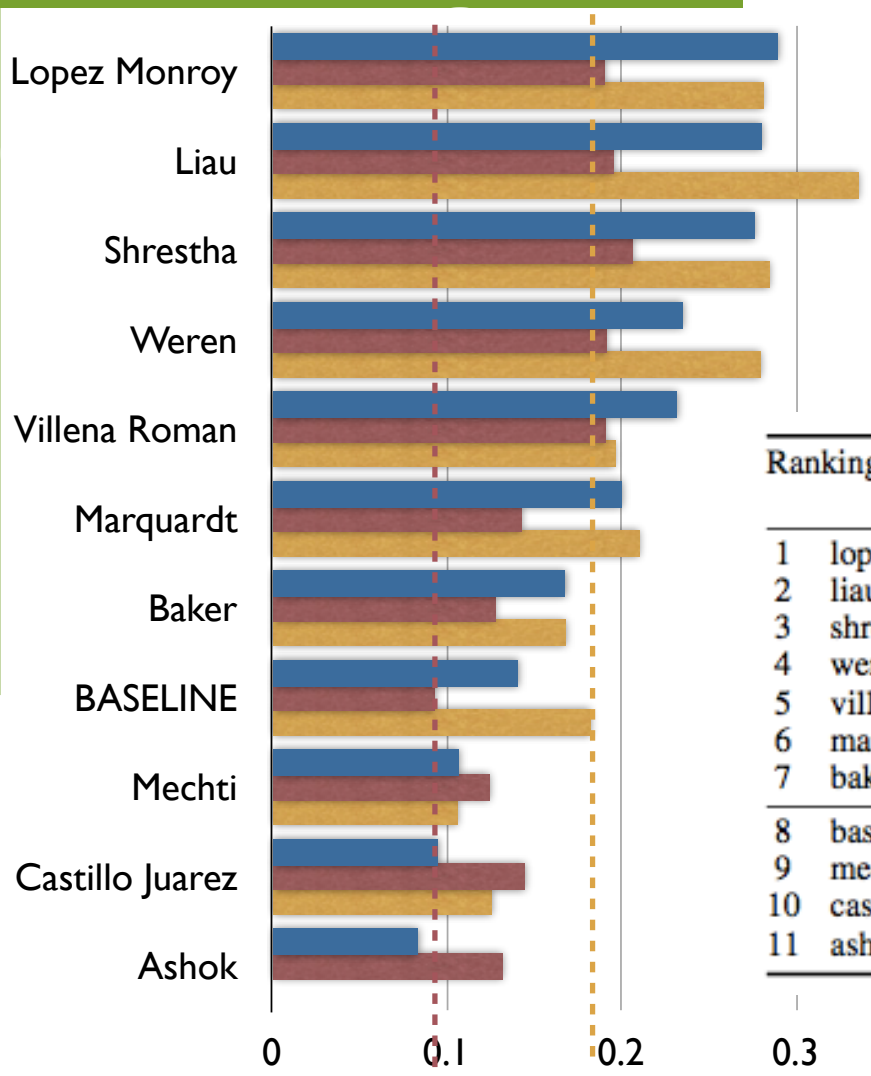
Ranking	Team	Average	Social Media		Blogs		Twitter		Reviews
			EN	ES	EN	ES	EN	ES	EN
1	lopezmonroy14	0.2895	0.1902	0.2809	0.3077	0.3214	0.3571	0.3444	0.2247
2	liau14	0.2802	0.1952	0.3357	0.2692	0.2321	0.3506	0.3222	0.2564
3	shrestha14	0.2760	0.2062	0.2845	0.2308	0.2500	0.3052	0.4333	0.2223
4	weren14	0.2349	0.1914	0.2792	0.2949	0.1786	0.2013	0.2778	0.2211
5	villenaroman14	0.2315	0.1905	0.1961	0.3077	0.2321	0.2078	0.2667	0.2199
6	marquardt14	0.1998	0.1428	0.2102	0.1282	0.2679	0.1948	0.3111	0.1437
7	baker14	0.1677	0.1277	0.1678	0.1282	0.2321	0.1688	0.2111	0.1382
8	baseline	0.1404	0.0930	0.1820	0.0897	0.0536	0.1494	0.2333	0.1821
9	mechti14	0.1067	0.1244	0.1060	0.0897	0.1786	0.0584	0.1444	0.0451
10	castillojuarez14	0.0946	0.1445	0.1254	0.1795	0.0893	-	-	0.1236
11	ashok14	0.0834	0.1318	-	0.1282	-	0.1948	-	0.1291

All results below 30%

BASELINE: 14%

3 teams below baseline

Average results in Social Media



Most results better for ES than EN

The highest (ES) ~ 33.57%

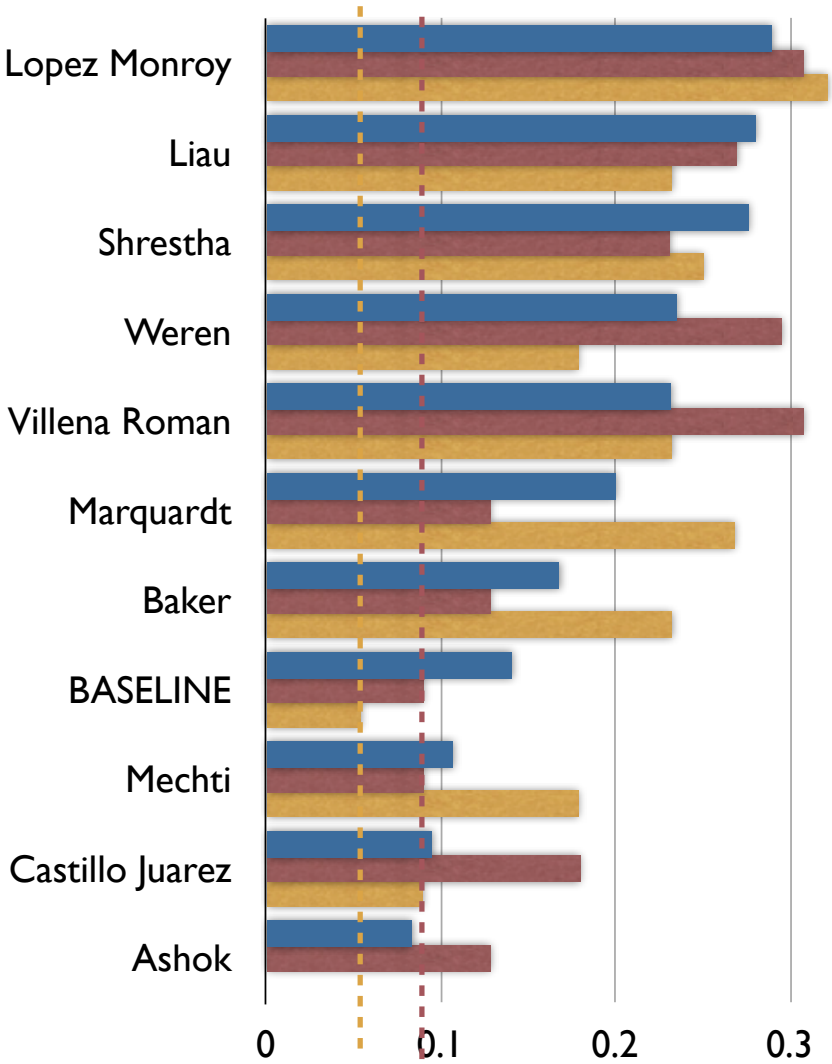
Most EN results lower than avg

English: All teams over baseline

Spanish: 3 teams below baseline

Ranking	Team	Average	Social Media		Blogs		Twitter		Reviews
			EN	ES	EN	ES	EN	ES	EN
1	lopezmonroy14	0.2895	0.1902	0.2809	0.3077	0.3214	0.3571	0.3444	0.2247
2	liau14	0.2802	0.1952	0.3357	0.2692	0.2321	0.3506	0.3222	0.2564
3	shrestha14	0.2760	0.2062	0.2845	0.2308	0.2500	0.3052	0.4333	0.2223
4	weren14	0.2349	0.1914	0.2792	0.2949	0.1786	0.2013	0.2778	0.2211
5	villenaroman14	0.2315	0.1905	0.1961	0.3077	0.2321	0.2078	0.2667	0.2199
6	marquardt14	0.1998	0.1428	0.2102	0.1282	0.2679	0.1948	0.3111	0.1437
7	baker14	0.1677	0.1277	0.1678	0.1282	0.2321	0.1688	0.2111	0.1382
8	baseline	0.1404	0.0930	0.1820	0.0897	0.0536	0.1494	0.2333	0.1821
9	mechti14	0.1067	0.1244	0.1060	0.0897	0.1786	0.0584	0.1444	0.0451
10	castillojuarez14	0.0946	0.1445	0.1254	0.1795	0.0893	-	-	0.1236
11	ashok14	0.0834	0.1318	-	0.1282	-	0.1948	-	0.1291

Average results in Blogs



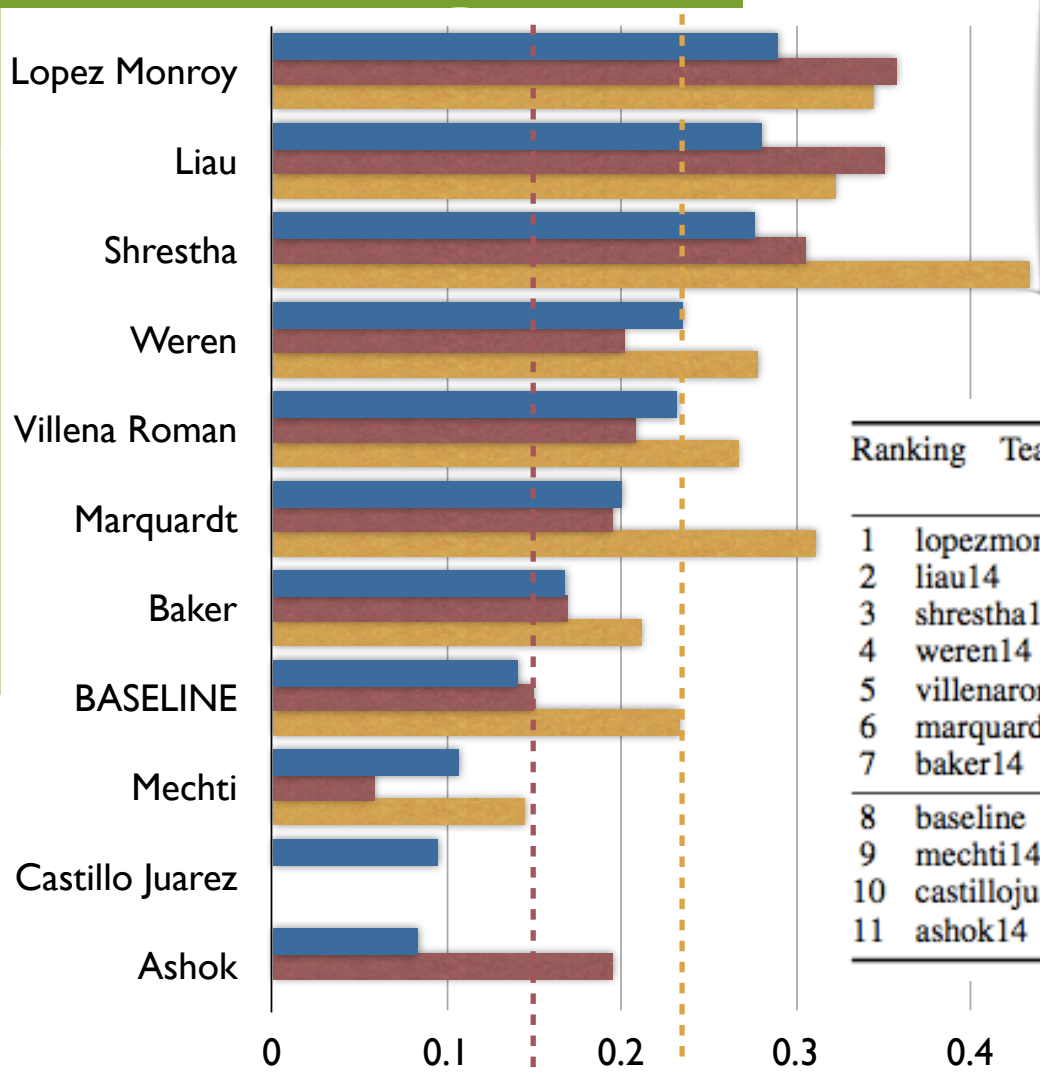
The highest result in Spanish ~ 32.14%

English: All teams over baseline (1=)

Spanish: All teams over baseline

Ranking	Team	Average	Social Media		Blogs		Twitter		Reviews
			EN	ES	EN	ES	EN	ES	EN
1	lopezmonroy14	0.2895	0.1902	0.2809	0.3077	0.3214	0.3571	0.3444	0.2247
2	liau14	0.2802	0.1952	0.3357	0.2692	0.2321	0.3506	0.3222	0.2564
3	shrestha14	0.2760	0.2062	0.2845	0.2308	0.2500	0.3052	0.4333	0.2223
4	weren14	0.2349	0.1914	0.2792	0.2949	0.1786	0.2013	0.2778	0.2211
5	villenaroman14	0.2315	0.1905	0.1961	0.3077	0.2321	0.2078	0.2667	0.2199
6	marquardt14	0.1998	0.1428	0.2102	0.1282	0.2679	0.1948	0.3111	0.1437
7	baker14	0.1677	0.1277	0.1678	0.1282	0.2321	0.1688	0.2111	0.1382
8	baseline	0.1404	0.0930	0.1820	0.0897	0.0536	0.1494	0.2333	0.1821
9	mechti14	0.1067	0.1244	0.1060	0.0897	0.1786	0.0584	0.1444	0.0451
10	castillojuarez14	0.0946	0.1445	0.1254	0.1795	0.0893	-	-	0.1236
11	ashok14	0.0834	0.1318	-	0.1282	-	0.1948	-	0.1291

Average results in Twitter



The highest result in Spanish ~ 43.33%

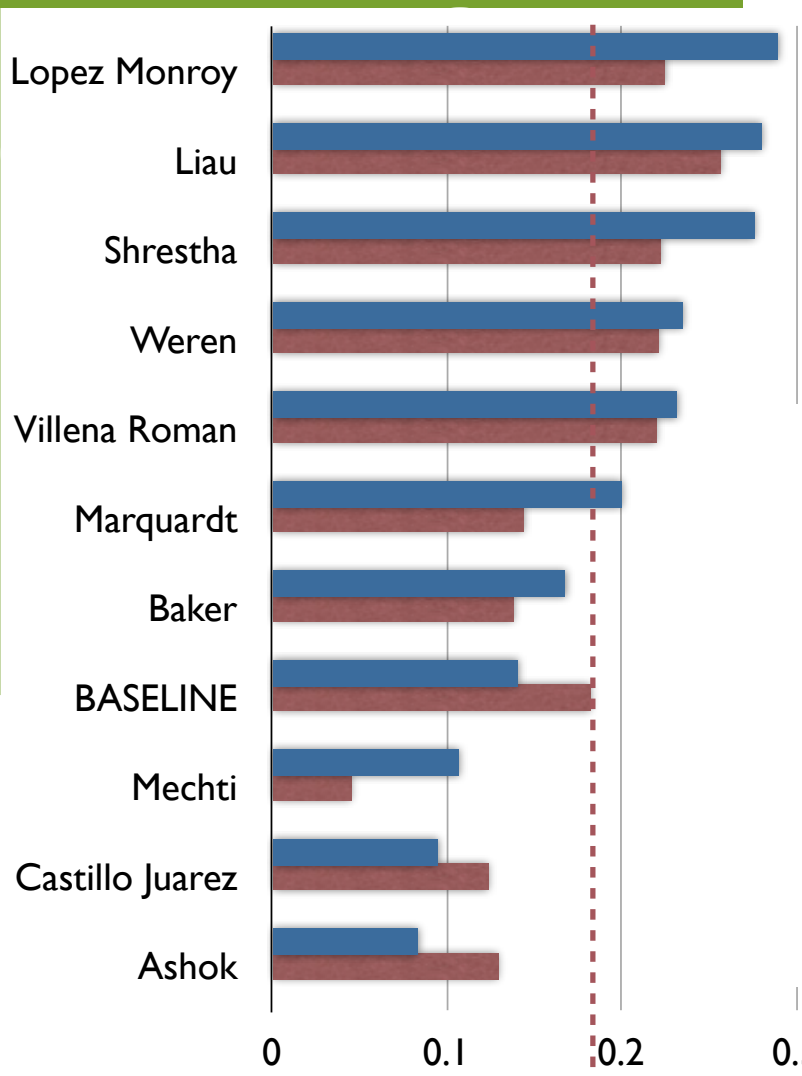
Most results higher than avg.

English: 1 team below baseline

Spanish: 2 teams below baseline

Ranking	Team	Average	Social Media		Blogs		Twitter		Reviews
			EN	ES	EN	ES	EN	ES	EN
1	lopezmonroy14	0.2895	0.1902	0.2809	0.3077	0.3214	0.3571	0.3444	0.2247
2	liau14	0.2802	0.1952	0.3357	0.2692	0.2321	0.3506	0.3222	0.2564
3	shrestha14	0.2760	0.2062	0.2845	0.2308	0.2500	0.3052	0.4333	0.2223
4	weren14	0.2349	0.1914	0.2792	0.2949	0.1786	0.2013	0.2778	0.2211
5	villenaroman14	0.2315	0.1905	0.1961	0.3077	0.2321	0.2078	0.2667	0.2199
6	marquardt14	0.1998	0.1428	0.2102	0.1282	0.2679	0.1948	0.3111	0.1437
7	baker14	0.1677	0.1277	0.1678	0.1282	0.2321	0.1688	0.2111	0.1382
8	baseline	0.1404	0.0930	0.1820	0.0897	0.0536	0.1494	0.2333	0.1821
9	mechti14	0.1067	0.1244	0.1060	0.0897	0.1786	0.0584	0.1444	0.0451
10	castillojuarez14	0.0946	0.1445	0.1254	0.1795	0.0893	-	-	0.1236
11	ashok14	0.0834	0.1318	-	0.1282	-	0.1948	-	0.1291

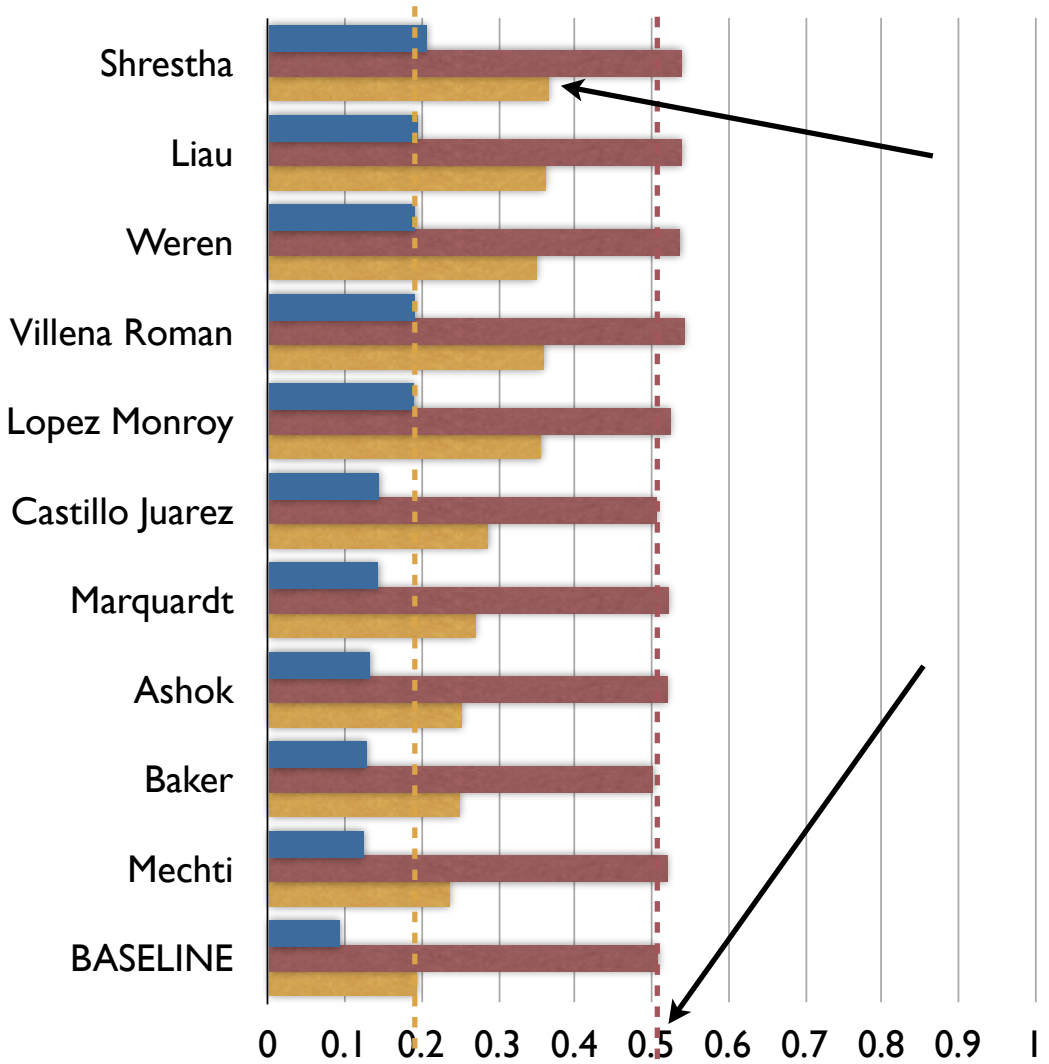
Average results in Reviews



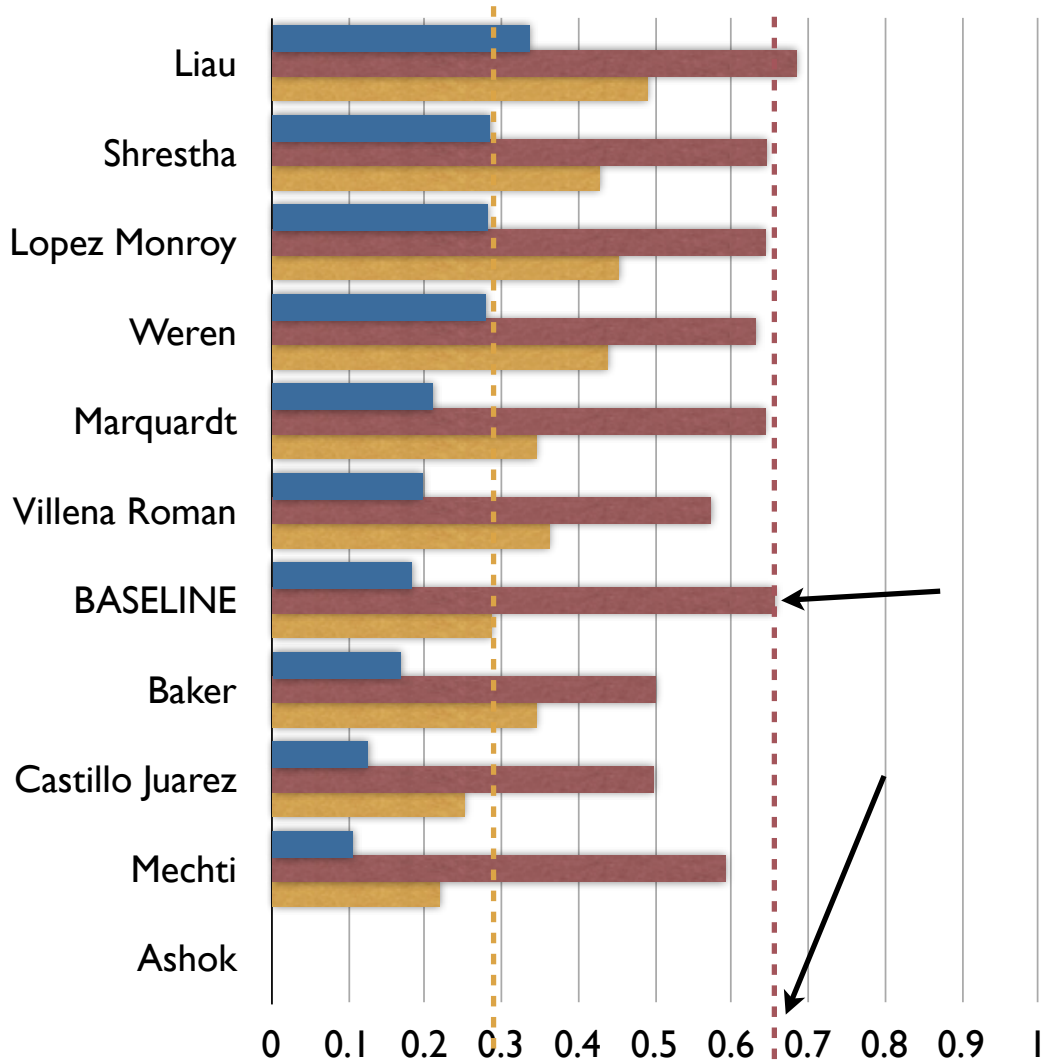
The highest result ~ 25.64%
Most results lower than avg.
5 teams below baseline

Ranking	Team	Average	Social Media		Blogs		Twitter		Reviews
			EN	ES	EN	ES	EN	ES	EN
1	lopezmonroy14	0.2895	0.1902	0.2809	0.3077	0.3214	0.3571	0.3444	0.2247
2	liau14	0.2802	0.1952	0.3357	0.2692	0.2321	0.3506	0.3222	0.2564
3	shrestha14	0.2760	0.2062	0.2845	0.2308	0.2500	0.3052	0.4333	0.2223
4	weren14	0.2349	0.1914	0.2792	0.2949	0.1786	0.2013	0.2778	0.2211
5	villenaroman14	0.2315	0.1905	0.1961	0.3077	0.2321	0.2078	0.2667	0.2199
6	marquardt14	0.1998	0.1428	0.2102	0.1282	0.2679	0.1948	0.3111	0.1437
7	baker14	0.1677	0.1277	0.1678	0.1282	0.2321	0.1688	0.2111	0.1382
8	baseline	0.1404	0.0930	0.1820	0.0897	0.0536	0.1494	0.2333	0.1821
9	mechti14	0.1067	0.1244	0.1060	0.0897	0.1786	0.0584	0.1444	0.0451
10	castillojuarez14	0.0946	0.1445	0.1254	0.1795	0.0893	-	-	0.1236
11	ashok14	0.0834	0.1318	-	0.1282	-	0.1948	-	0.1291

Results in Social Media

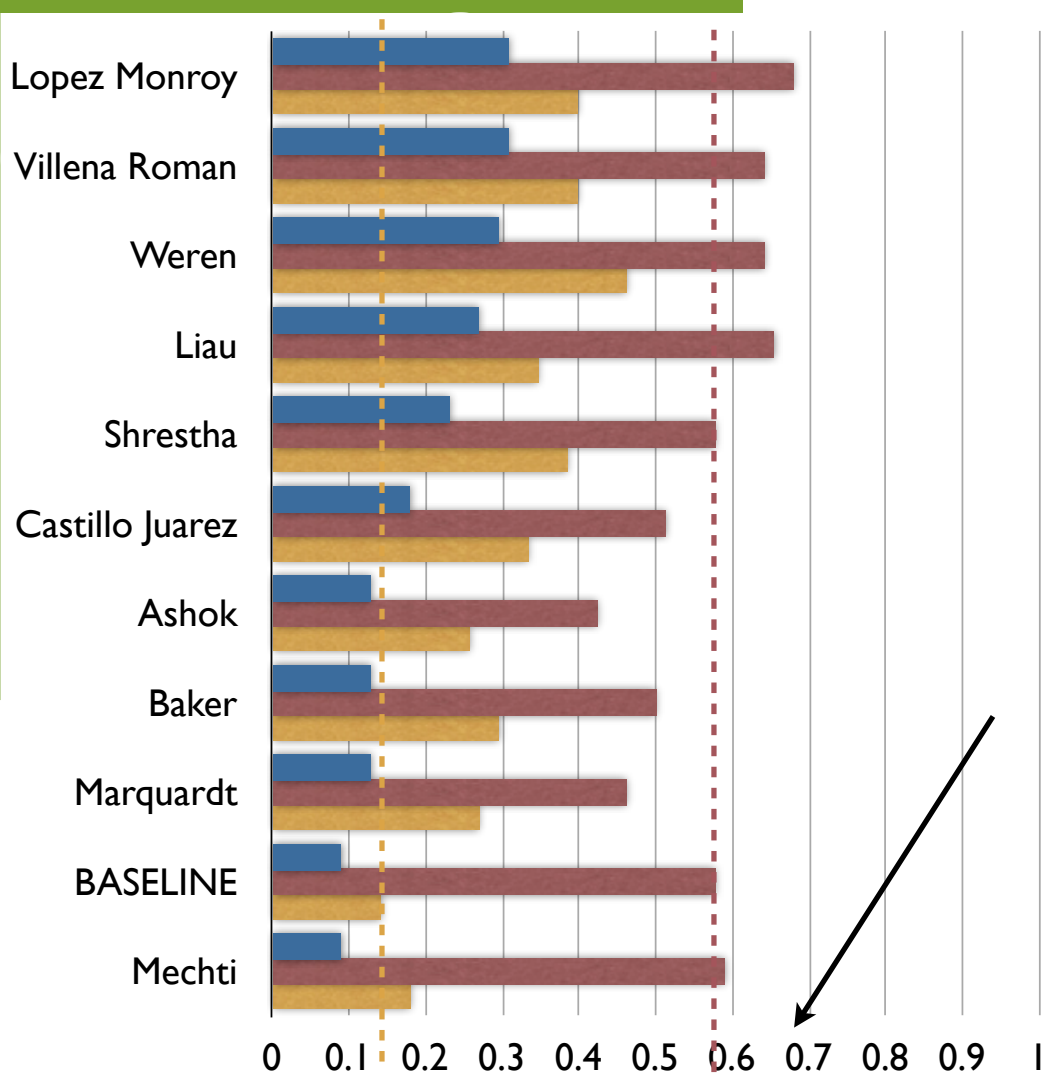


EN (joint, gender, age)

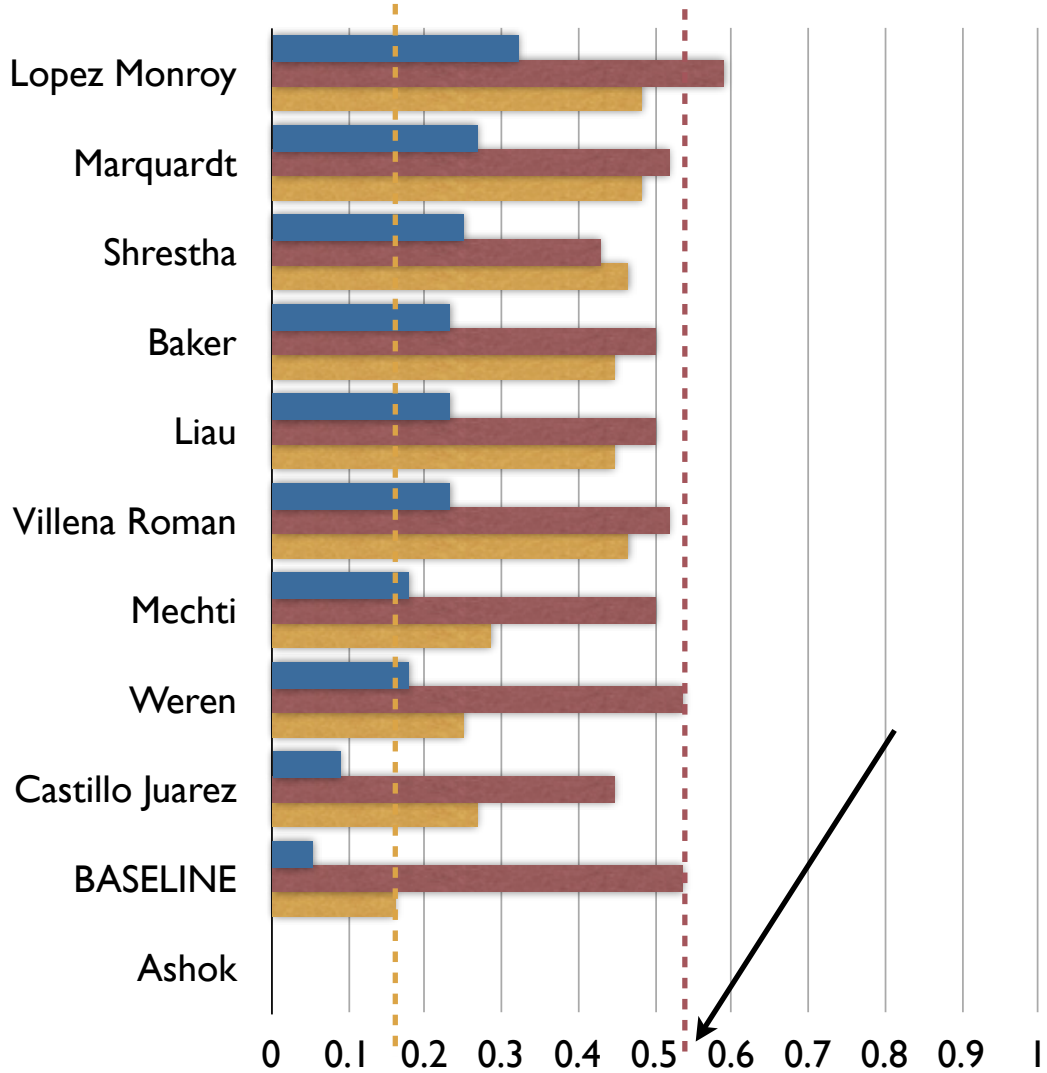


ES (joint, gender, age)

Results in Blogs

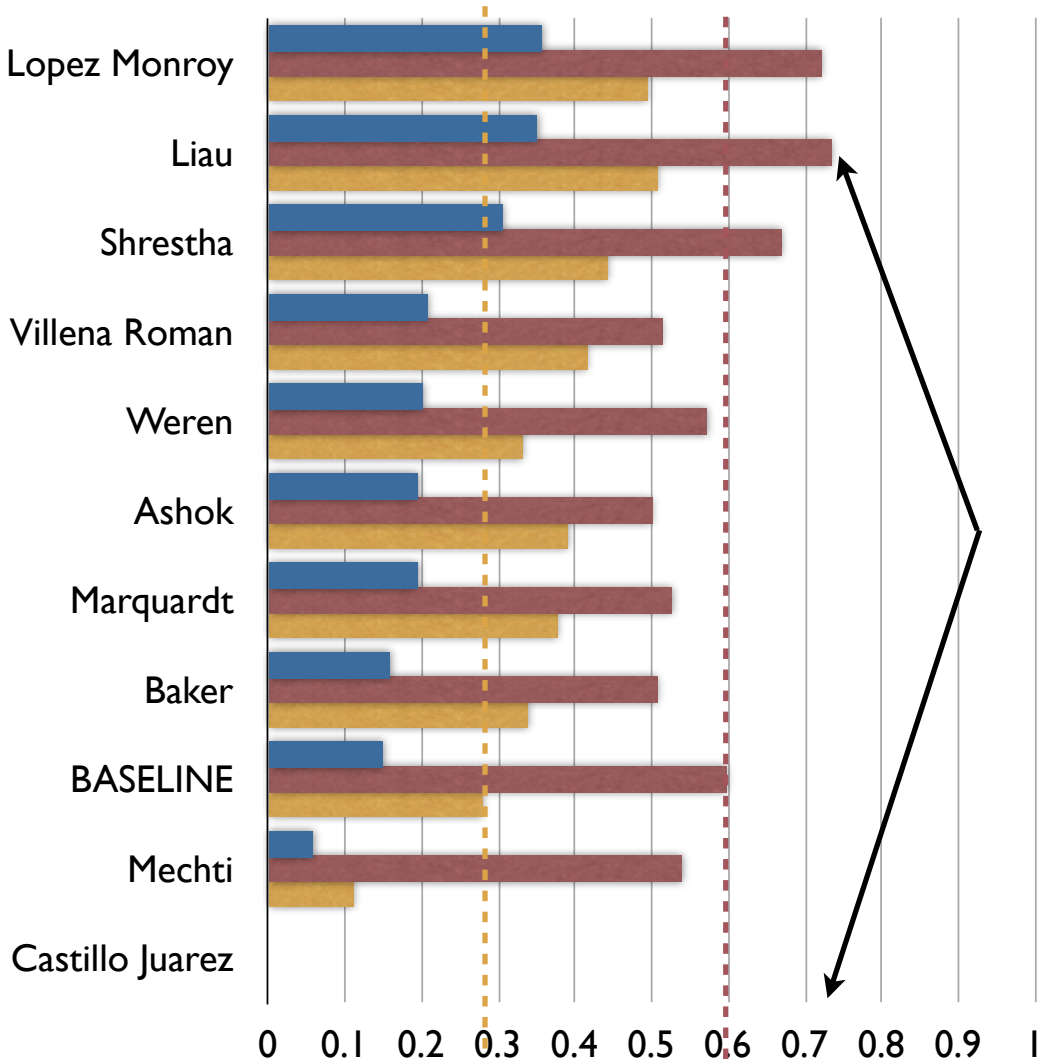


EN (joint, gender, age)

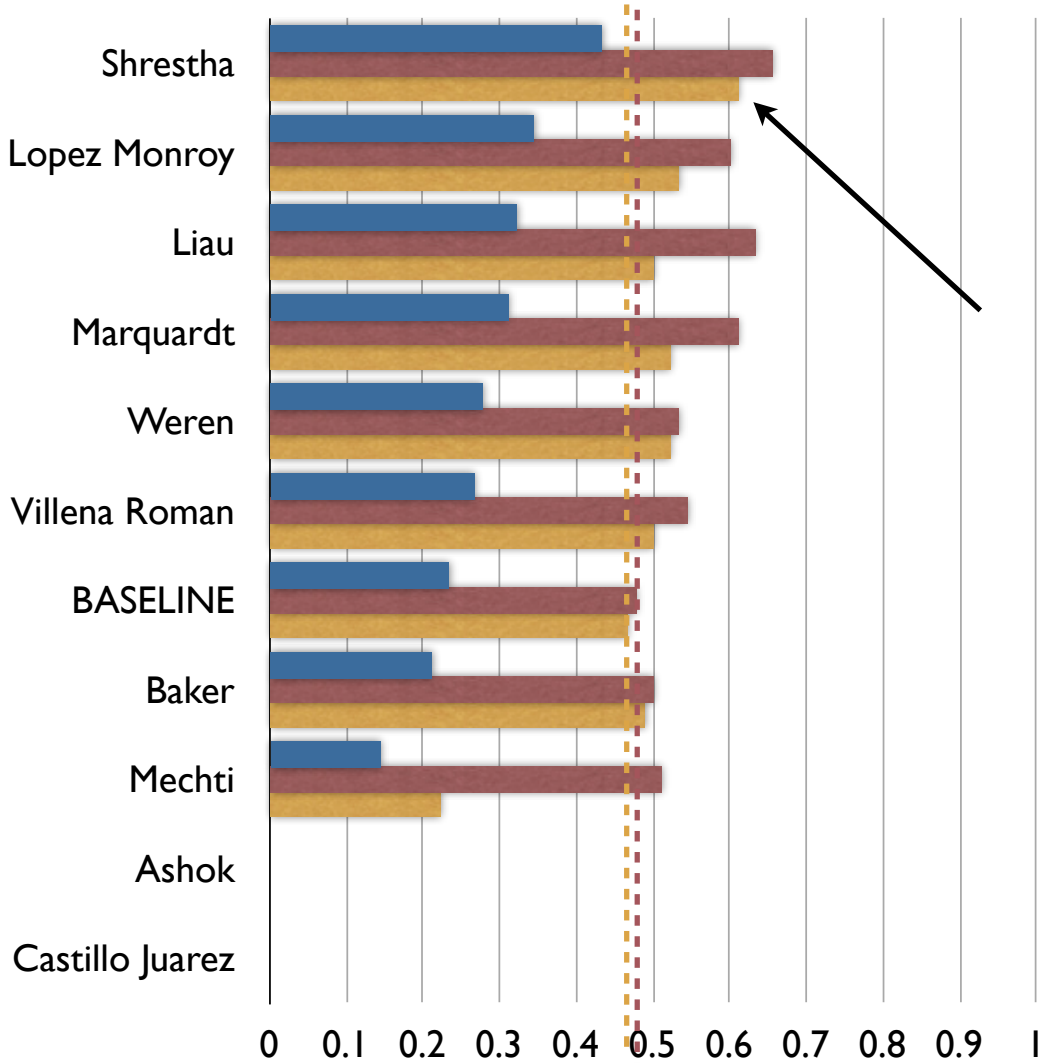


ES (joint, gender, age)

Results in Twitter

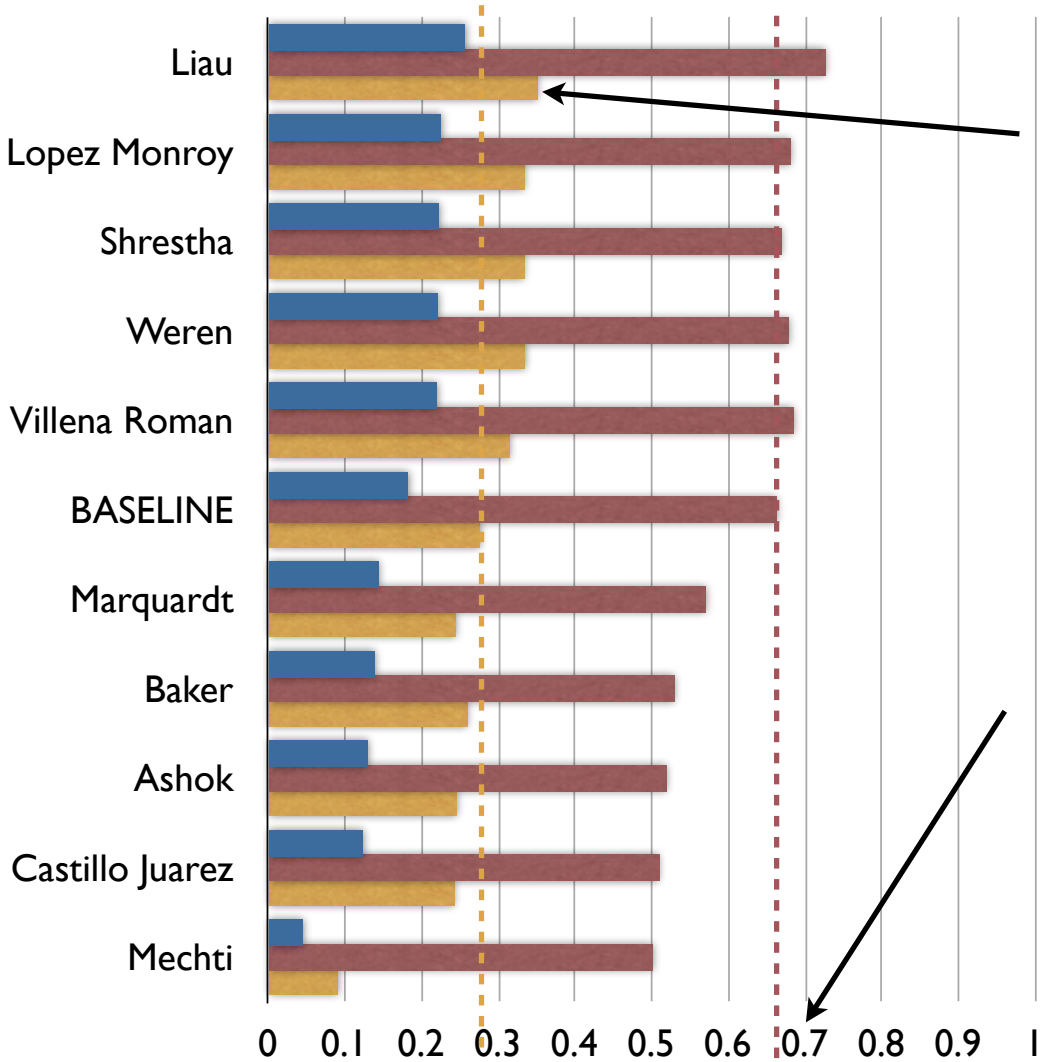


EN (joint, gender, age)



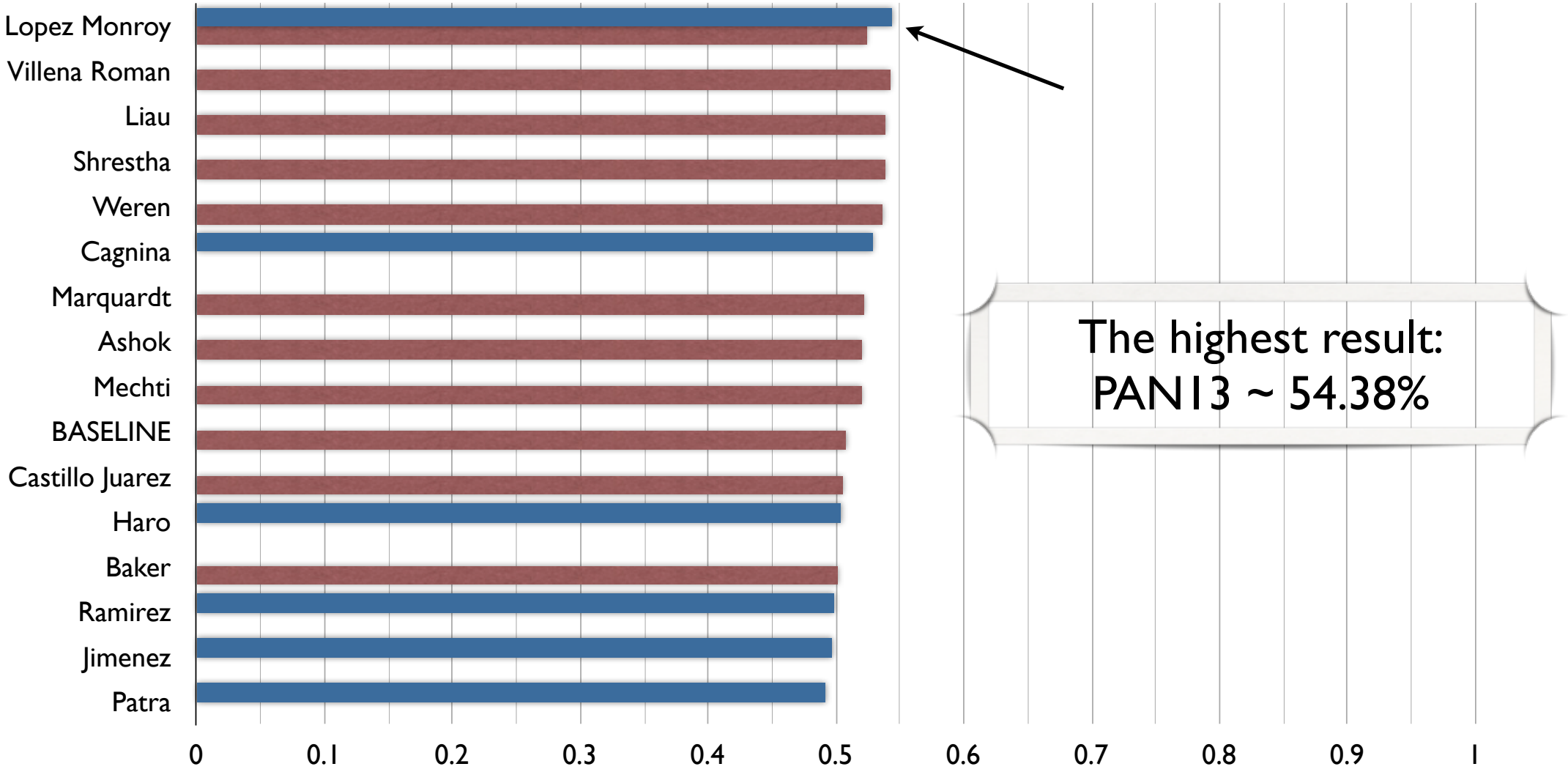
ES (joint, gender, age)

Results in Reviews

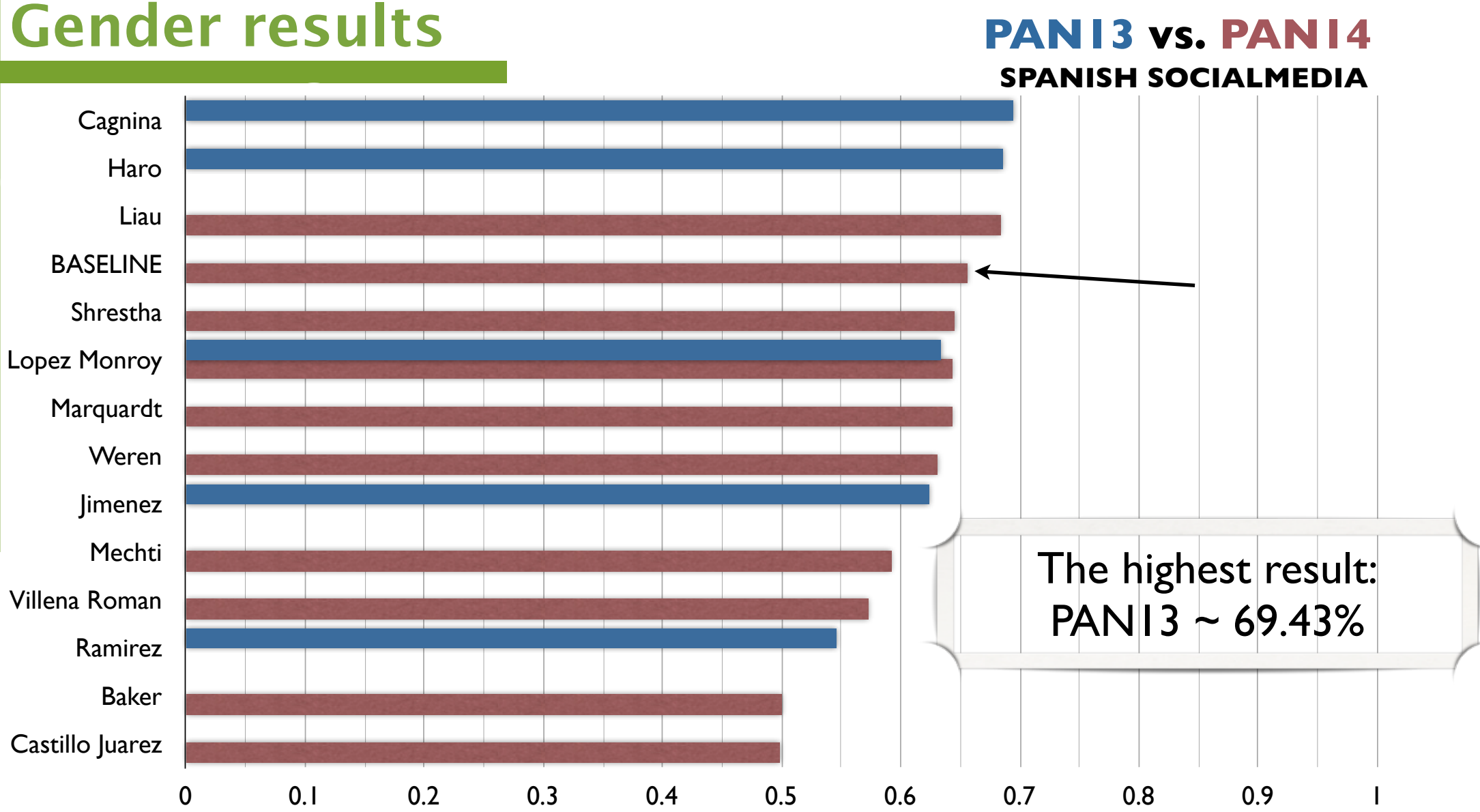


EN (joint, gender, age)

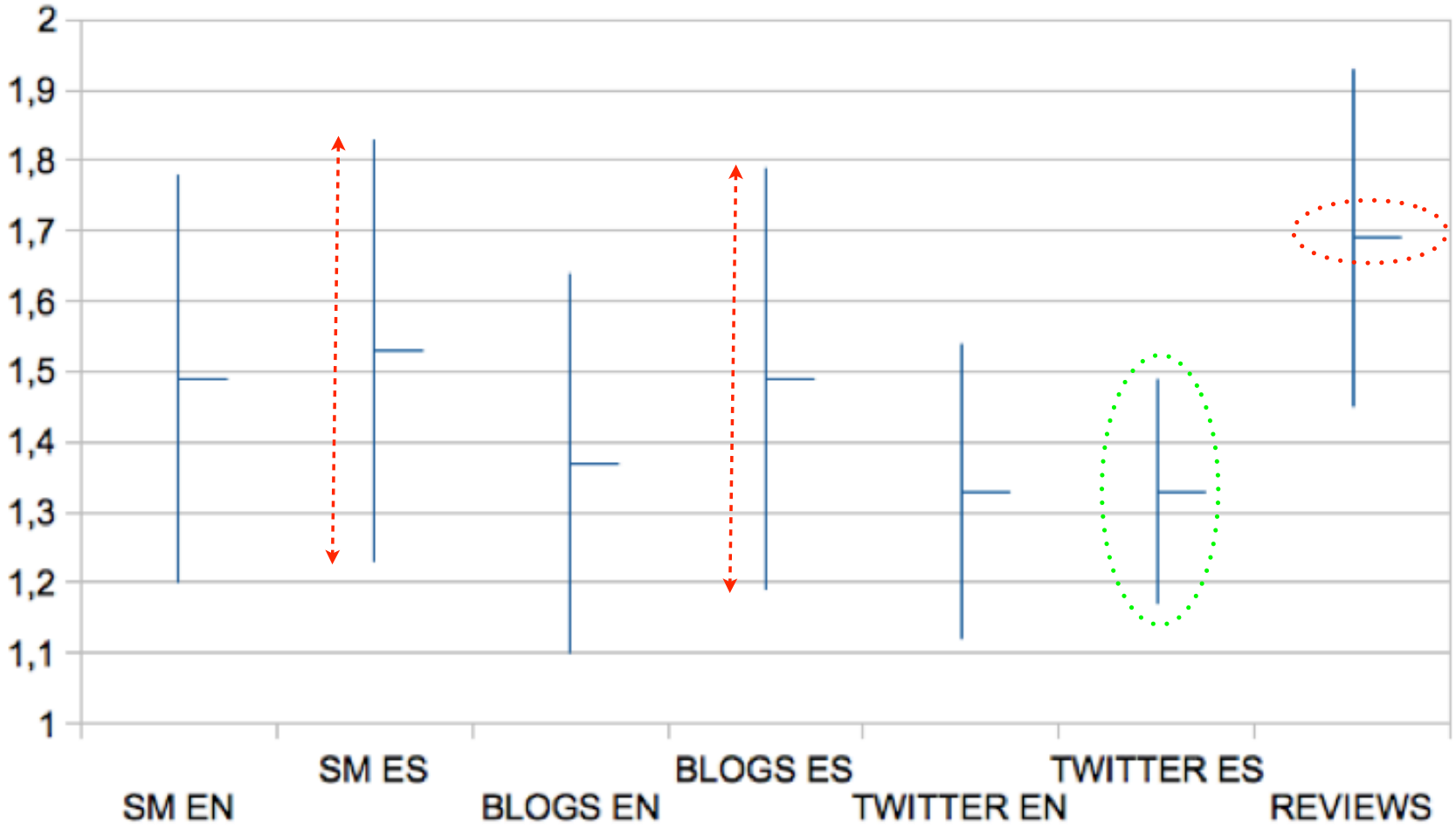
Gender results



Gender results



Distances in misclassified age



Conclusions

- The highest accuracies were achieved in Twitter
  - Higher number of documents per profile
  - More spontaneous language
- The lowest accuracies were achieved in English social media and hotel reviews
- The highest distance between predicted and truth classes in age identification occur in hotel reviews
  - A further analysis is needed to understand if there are cases of deceptive opinions

Industry at PAN (Author Profiling)

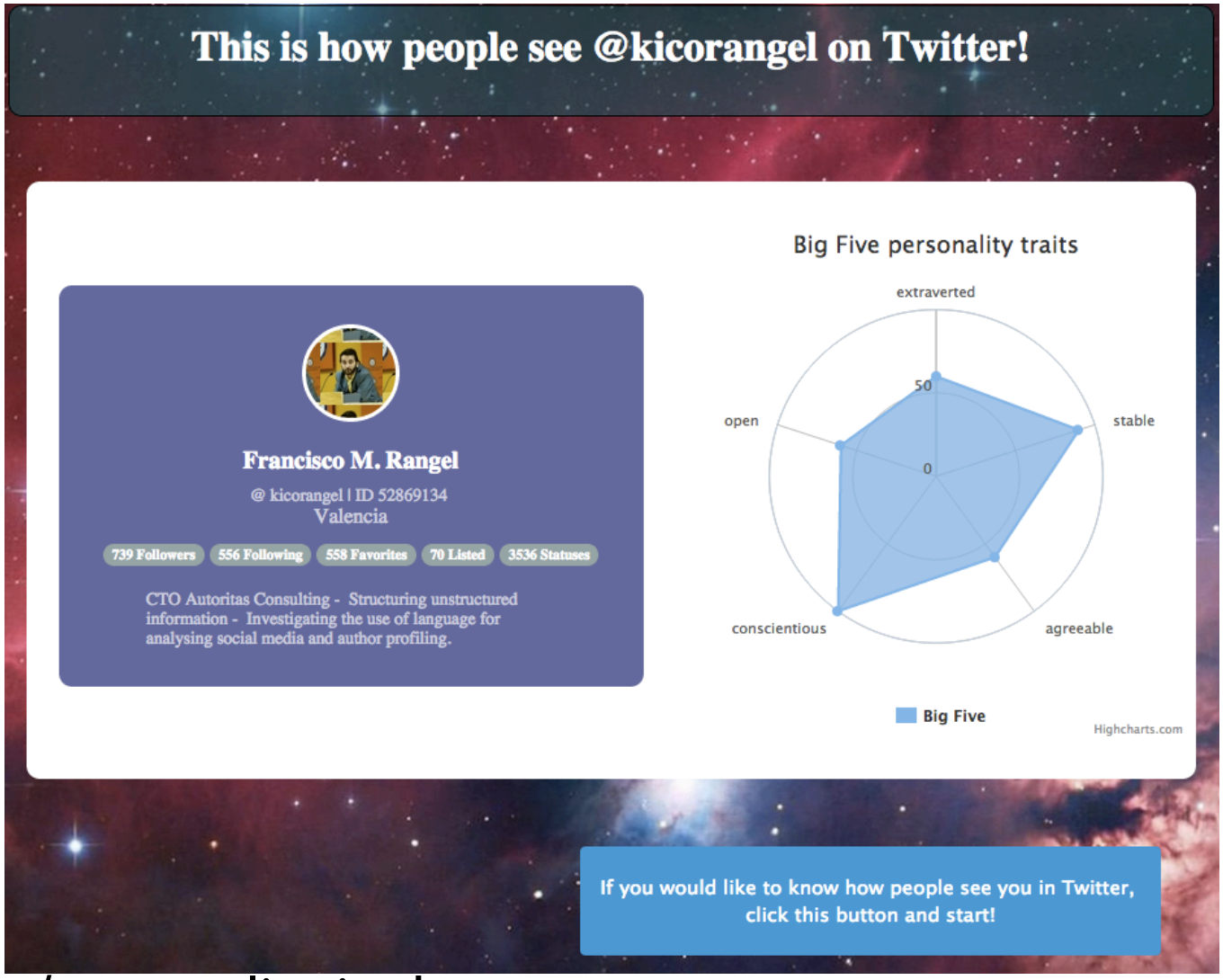
Organisers	autoritas® nuevas ideas, nuevas soluciones
Collaborators	LLORENTE & CUENCA
Sponsors	COREX BUILDING KNOWLEDGE SOLUTIONS
Participants	DAEDALUS

Next year...

- AGE + GENDER

+

PERSONALITY RECOGNITION!



http://personality.altervista.org/personalitwit.php








Francisco Rangel

autoritas
nuevas ideas, nuevas soluciones

Paolo Rosso

Irina Chugur

Martin Potthast

Benno Stein

Martin Trenkmann

Ben Verhoeven

Walter Daelemans

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNED

Bauhaus-Universität Weimar

University of Antwerp

On behalf of the AP task organisers:
Thank you very much for participating!
We hope to see you again next year!