Author Profiling

PAN-AP-2015 - CLEF 2015

Toulouse, 8-11 September 2015

PAN

Francisco Rangel
Autoritas / Universitat Politècnica de València

Fabio Celli
University of Trento

Paolo Rosso
Universitat Politècnica de València

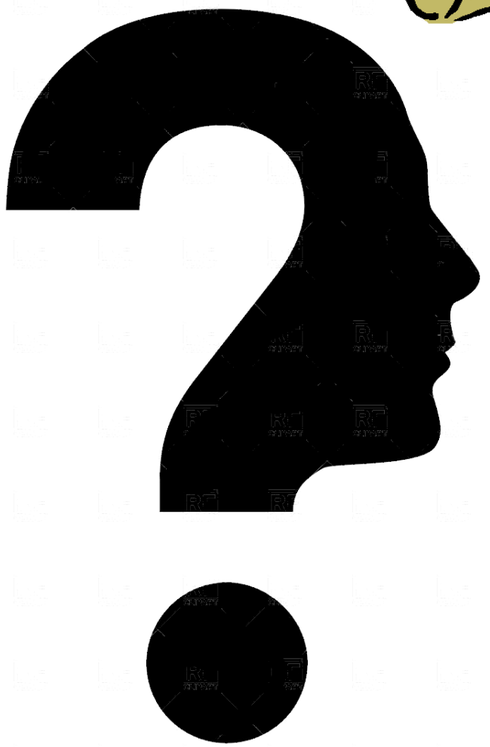
Martin Potthast, Benno Stein
Bauhaus-Universität Weimar

Walter Daelemans
University of Anwerp

What's Author Profiling?

Gender?

Personality traits?



Age?

Emotions?

Native language?

Language variety?

Author Profile... Who is who?

Why Author Profiling?

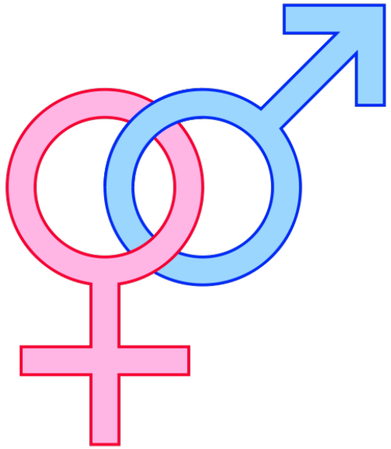
Forensics	Security	Marketing
Language as evidence	***Profile possible delinquents***	***Segmenting users***

Task Goal

To identify age and gender

+

Personality Traits

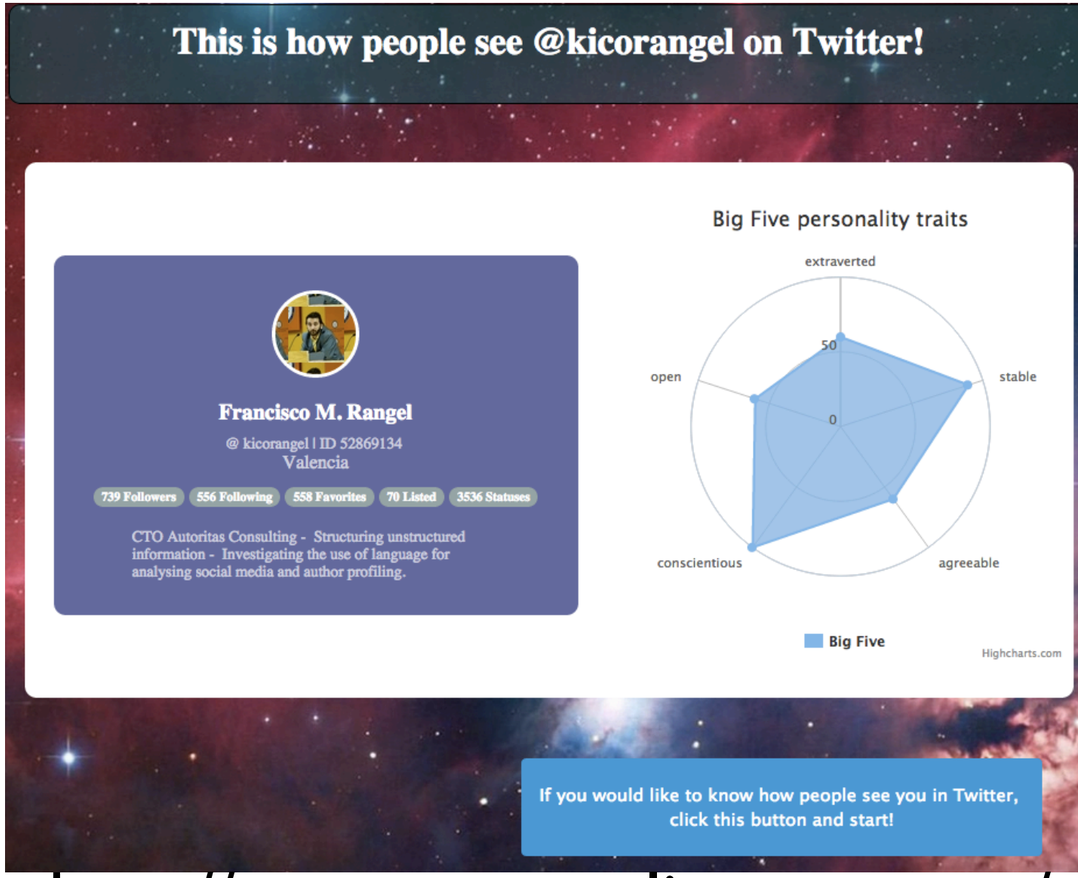


Languages

- English
- Spanish
- Italian
- Dutch

Data collection and labeling

- We provided an online test to collect Twitter users:
  - Age and gender were provided by users.
  - Personality traits were self-assessed with the BFI-10 online test.



http://your-personality-test.com/

- It was hard to make users respond to the test:
  - We launched advertisments campaigns.
  - We had a high number of visits to our online test.
  - But only few people answer the test.

Corpus

	Training				Early birds				Test			
	EN	ES	IT	DU	EN	ES	IT	DU	EN	ES	IT	DU
Users	152	110	38	34	42	30	12	10	142	88	36	32
18-24	58	22			16	6			56	18		
25-34	60	56			16	14			58	44		
35-49	22	22			6	6			20	18		
50+	12	10			4	4			8	8		
Male	76	55	19	17	21	15	6	5	71	44	18	16
Female	76	55	19	17	21	15	6	5	71	44	18	16
E (mean)	0.16	0.18	0.17	0.24	0.19	0.15	0.16	0.21	0.17	0.16	0.15	0.24
S (mean)	0.14	0.07	0.20	0.21	0.11	0.07	0.24	0.23	0.13	0.09	0.20	0.22
A (mean)	0.12	0.14	0.22	0.13	0.14	0.17	0.17	0.14	0.14	0.14	0.19	0.15
C (mean)	0.17	0.24	0.18	0.14	0.17	0.22	0.22	0.17	0.17	0.21	0.21	0.17
O (mean)	0.24	0.18	0.23	0.29	0.28	0.19	0.29	0.27	0.26	0.19	0.25	0.28

Evaluation measures

Age and Gender

Personality

Accuracy for Gender

Accuracy for Age

RMSE

$$\sqrt{\frac{\sum_{i=1}^{n}(\hat{f}_i - f_i)^2}{n}}$$

Joint Accuracy

$$rank = \frac{(1 - RMSE) + joint\_accuracy}{2}$$

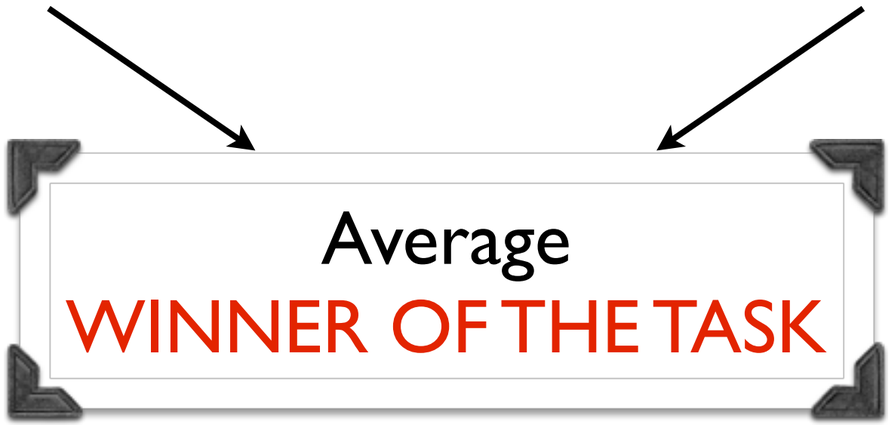
Participants' ranking

Rank for English

Rank for Spanish

Rank for Italian

Rank for Dutch



Statistical significance

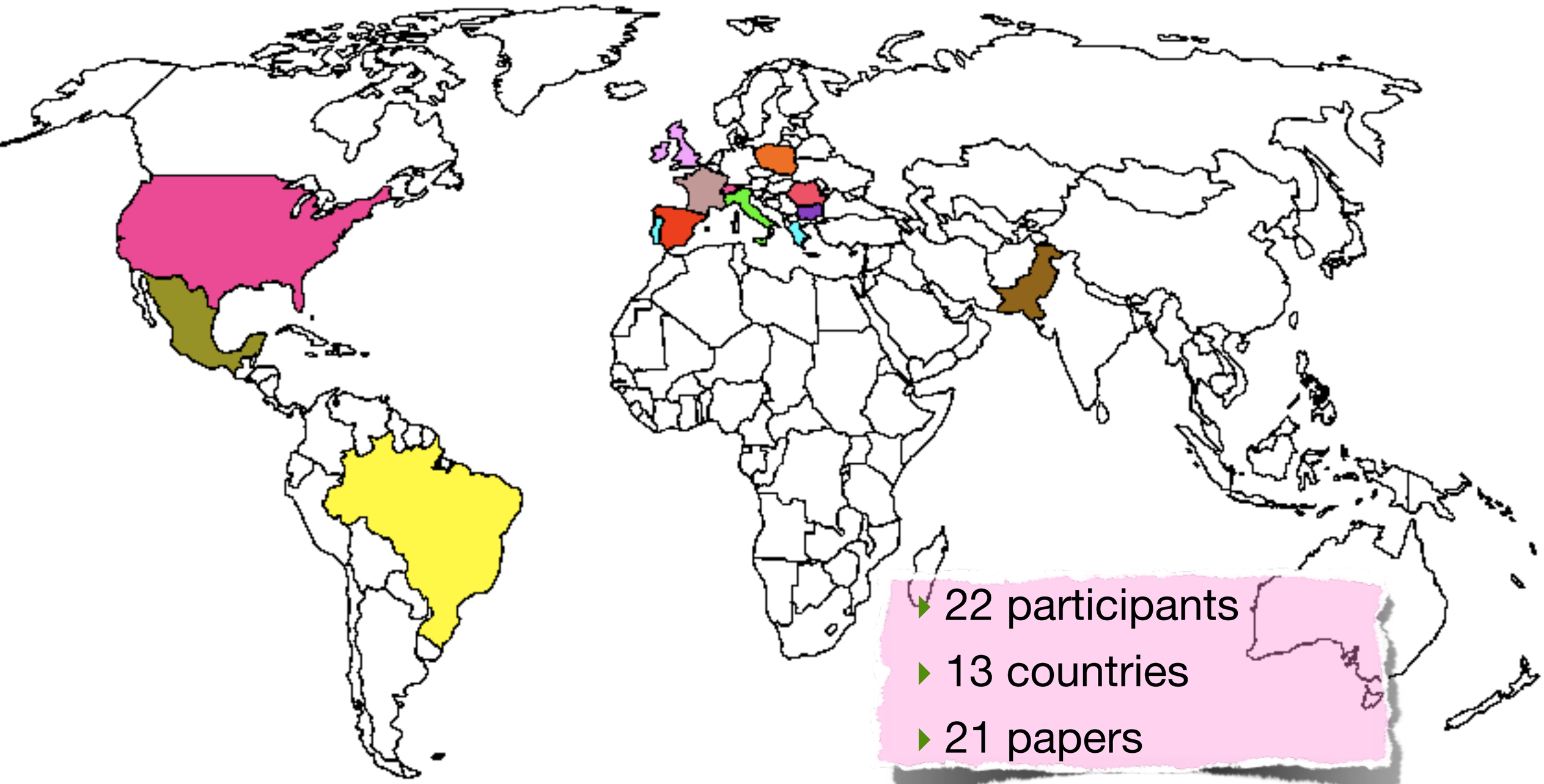
Approximate randomisation testing*

*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

Participants



- 22 participants
- 13 countries
- 21 papers

Approaches

- What kind of ...

Preprocessing

Features

Methods

... did the teams perform?

Approaches

Preprocessing

HTML Cleaning to obtain plain text	[arroju][grivas][cheema][ashraf]
Hashtags, urls, mentions	[arroju][gonzalezgallardo][grivas][maharjan][nowson]
Remove RT and shares	[bartoli][poulston]
Lowercased, remove numbers and stopwords	[weren]
Emojis	[nowson]
Remove tweets <5 words	[markov]
SVM for Feature Selection	[miculicich]
Recursive Feature Extraction	[guyon]

Approaches

Features

punctuation signs	[miculicich][mezaruiz][ameer]
emoticons	[nowson] [mezaruiz] [markov][teisseyre]
word length	[grivas]
sentence length	[ameer]
character flooding	[gimenez][kiprov][nowson]
verbosity	[sulea]
letter case	[gimenez][grivas] [kiprov]
question marks, question sentences	[maharjan] [ameer]

Approaches

Features

Specific Twitter elements: hashtags, links, mentions	[gimenez][grivas] [kiprov][miculicich] [nowson] [mezaruiz] [markov]
Latent Semantic Analysis	[maharjan][mccollister][miculicich] [poulston][ashraf]
Family Tokens (my wife/husband, my girlfriend/boyfriend...)	[maharjan]
Most discriminant words among classes	[cheema]
Named Entities	[nowson]

Approaches

Features

character n-grams	[gonzalezgallardo] [maharjan][sulea]
word n-grams	[arroju] [gimenez][cheema] [nowson] [mezaruiz]
tf-idf n-grams	[gimenez][grivas] [mezaruiz] [sulea]
POS n-grams	[gonzalezgallardo] [mezaruiz]
lemmas, words, relations, POS... n-grams	[markov]

Approaches

Features

LIWC	[arroju] [bartoli][miculicich]
NRC	[gimenez][kiprov]
Polarity, emotions	[gimenez] [kiprov][nowson] [mezaruiz] [teisseyre]
Irony, Taboo	[mezaruiz]
IR features: cosine similarity, Okapi BM25	[weren]
LDA + Second order representation	[alvarezcarmona]

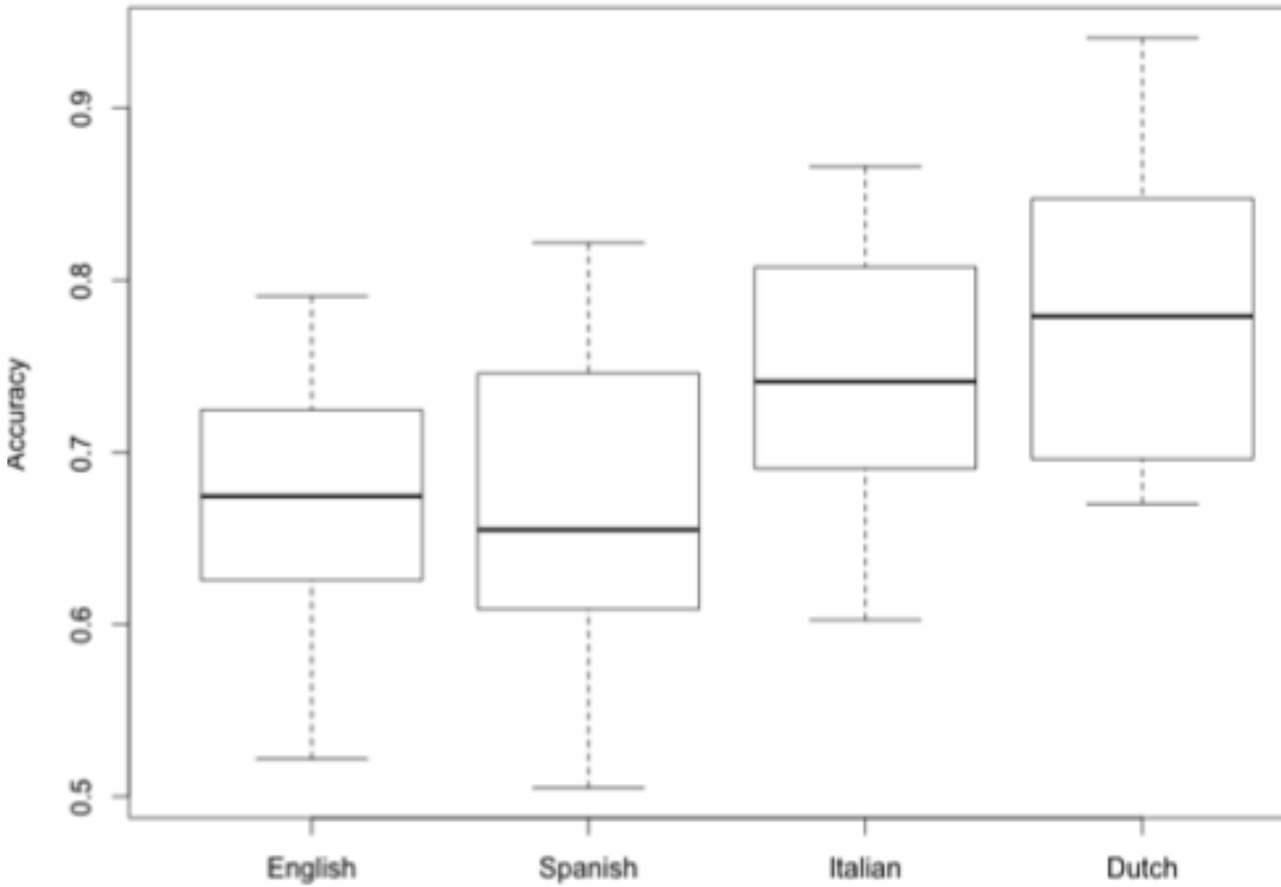
Approaches

Methods

Age and Gender	Support Vector Machines	1 [gonzalezgallardo][cheema][markov][grivas] [kiprov][nowson] [poulston][bartoli]
	Random Forest	[mezaruiz][ashraf]
	Rotation Forest	[mccollister]
Personality Traits	Support Vector Machines	1 [gonzalezgallardo][cheema] [markov]
	Random Forest	[bartoli]
	J48	[mezaruiz][ashraf]
	Rotation Forest	[mccollister]
	Linear Discriminant Analysis	[miculicich]
	SGD	[arroju]
	Linear Regression	[gimenez]
	Ridge for Regression	[sulea]
	Logic Regression	[maharjan]

Global ranking per language

Ranking	Team	Global	English	Spanish	Italian	Dutch
1	alvarezcarmona15	**0.8404**	**0.7906**	**0.8215**	0.8089	**0.9406**
2	gonzalesgallardo15	0.8346	0.7740	0.7745	**0.8658**	0.9242
3	grivas15	0.8078	0.7487	0.7471	0.8295	0.9058
4	kocher15	0.7875	0.7037	0.7735	0.8260	0.8469
5	sulea15	0.7755	0.7378	0.7496	0.7509	0.8637
6	miculicich15	0.7584	0.7115	0.7302	0.7442	0.8475
7	nowson15	0.7338	0.6039	0.6644	0.8270	0.8399
8	weren15	0.7223	0.6856	0.7449	0.7051	0.7536
9	poulston15	0.7130	0.6743	0.6918	0.8061	0.6796
10	maharjan15	0.7061	0.6623	0.6547	0.7411	0.7662
11	mccollister15	0.6960	0.6746	0.5727	0.7015	0.8353
12	arroju15	0.6875	0.6996	0.6535	0.7126	0.6843
13	gimenez15	0.6857	0.5917	0.6129	0.7590	0.7790
14	bartoli15	0.6809	0.6557	0.5867	0.6797	0.8016
15	ameer15	0.6685	0.6379	0.6044	0.7055	0.7260
16	cheema15	0.6495	0.6130	0.6353	0.6774	0.6723
17	teisseyre15	0.6401	0.7489	0.5049	0.6024	0.7042
18	mezaruiz15	0.6204	0.5217	0.6215	0.6682	0.6703
19	bayot15	0.6178	0.5253	0.5932	0.6644	0.6881
	ashraf15	-	0.5854	-	-	-
	kiprov15	-	0.7211	0.7889	-	-
	markov15	-	0.5890	0.5874	-	0.6798



Best results per language and task

	Age and Gender				Personality Traits				
Language	*Joint*	Gender	Age	*RMSE*	E	S	A	C	O
English	0.7254	0.8592	0.8380	0.1442	0.1250	0.1951	0.1305	0.1101	0.1198
Spanish	0.7727	0.9659	0.7955	0.1235	0.1319	0.1631	0.1034	0.1017	0.1108
Italian	-	0.8611	-	0.1044	0.0726	0.1555	0.0527	0.1093	0.0972
Dutch	-	0.9688	-	0.0563	0.0750	0.0637	0.0000	0.0619	0.0354

- Age and gender results much higher than in previous editions of PAN

English

	Age and Gender				Personality Traits				
Team	*Joint*	Gender	Age	*RMSE*	E	S	A	C	O
alvarezcarmona15	**0.7254**	**0.8592**	**0.8380**	**0.1442**	0.1278	0.2253	**0.1305**	0.1172	0.1202
ameer15	0.5070	0.6901	0.7183	0.2313	0.2131	0.3172	0.2154	0.1959	0.2149
arroju15	0.5704	0.7676	0.7042	0.1713	0.1636	0.2349	0.1513	0.1481	0.1584
ashraf15	0.3944	0.5563	0.6972	0.2236	0.2084	0.3151	0.1910	0.1897	0.2138
bartoli15	0.4718	0.6479	0.7465	0.1605	0.1480	0.2323	0.1360	0.1418	0.1445
bayot15	0.2465	0.5000	0.5915	0.1958	0.2137	0.2308	0.1634	0.1866	0.1844
cheema15	0.4225	0.5915	0.6690	0.1965	0.1878	0.2612	0.1766	0.1610	0.1959
gimenez15	0.3873	0.6338	0.5986	0.2039	0.1770	0.2781	0.1754	0.1819	0.2073
gonzalesgallardo15	0.6972	0.8521	0.7817	0.1491	0.1303	0.2151	0.1480	**0.1101**	0.1422
grivas15	0.6690	**0.8592**	0.7465	0.1716	0.1411	0.2039	0.1432	0.2249	0.1450
kiprov15	0.5915	0.8451	0.7254	0.1493	0.1416	0.2123	0.1411	0.1318	**0.1198**
kocher15	0.5563	0.7113	0.7113	0.1489	0.1417	0.2062	0.1427	0.1181	0.1358
maharjan15	0.5634	0.7465	0.6901	0.2388	0.2299	0.2647	0.2127	0.2222	0.2645
markov15	0.3662	0.5915	0.5845	0.1882	0.1806	0.2708	0.1570	0.1893	0.1434
mccollister15	0.5141	0.7254	0.7183	0.1649	0.1537	0.2205	0.1513	0.1443	0.1545
mezaruiz15	0.2183	0.5000	0.4085	0.1749	0.1676	0.2392	0.1572	0.1526	0.1582
miculicich15	0.5704	0.7887	0.6901	0.1475	**0.1250**	0.2247	0.1322	0.1330	0.1225
nowson15	0.3732	0.7746	0.4930	0.1655	0.1665	0.2059	0.1647	0.1483	0.1419
poulston15	0.5211	0.6901	0.7394	0.1725	0.1381	0.2223	0.1918	0.1749	0.1352
sulea15	0.6197	0.7676	0.7887	**0.1442**	0.1318	**0.1951**	0.1396	0.1297	0.1246
teisseyre15	0.6479	0.8310	0.7535	0.1500	0.1371	0.1990	0.1480	0.1309	0.1351
weren15	0.5563	0.7606	0.7042	0.1851	0.1597	0.2593	0.1768	0.1574	0.1722

Spanish

	Age and Gender				Personality Traits				
Team	*Joint*	Gender	Age	*RMSE*	E	S	A	C	O
alvarezcarmona15	**0.7727**	**0.9659**	**0.7955**	0.1297	**0.1319**	**0.1631**	0.1113	0.1168	0.1257
ameer15	0.4205	0.6932	0.5341	0.2116	0.2786	0.2806	0.1430	0.1410	0.2145
arroju15	0.4886	0.7500	0.6932	0.1817	0.1980	0.2125	0.1727	0.1785	0.1469
bartoli15	0.3295	0.8523	0.4205	0.1562	0.1701	0.1867	0.1463	0.1320	0.1459
bayot15	0.3636	0.6136	0.5682	0.1773	0.1853	0.2025	0.1593	0.1852	0.1540
cheema15	0.4545	0.8409	0.5682	0.1839	0.1599	0.2479	0.1880	0.1526	0.1712
gimenez15	0.4205	0.6250	0.5682	0.1947	0.2097	0.2440	0.1729	0.1853	0.1617
gonzalesgallardo15	0.7045	0.8977	0.7273	0.1555	0.1406	0.2094	0.1168	0.1709	0.1398
grivas15	0.6818	0.9432	0.6932	0.1876	0.1762	0.1965	0.1557	0.2745	0.1353
kiprov15	0.7273	0.9091	0.7841	0.1495	0.1625	0.1884	0.1249	0.1386	0.1334
kocher15	0.6705	0.8182	0.7386	**0.1235**	0.1373	0.1641	**0.1034**	**0.1017**	**0.1108**
maharjan15	0.5795	0.7955	0.6250	0.2702	0.3008	0.2880	0.2569	0.2357	0.2696
markov15	0.3864	0.6591	0.5114	0.2116	0.1877	0.2644	0.1916	0.2400	0.1742
mccollister15	0.3182	0.6818	0.5000	0.1728	0.1877	0.2098	0.1674	0.1588	0.1403
mezaruiz15	0.4091	0.8295	0.5114	0.1660	0.1729	0.2035	0.1536	0.1473	0.1530
miculicich15	0.6250	0.9205	0.6818	0.1647	0.1856	0.1971	0.1327	0.1402	0.1679
nowson15	0.4886	0.7727	0.6705	0.1598	0.1578	0.2023	0.1358	0.1461	0.1571
poulston15	0.5455	0.8409	0.5909	0.1619	0.1669	0.2285	0.1398	0.1412	0.1329
sulea15	0.6591	0.8750	0.7500	0.1599	0.1703	0.1816	0.1501	0.1559	0.1417
teisseyre15	0.2159	0.5568	0.3636	0.2060	0.1957	0.2446	0.1937	0.2194	0.1768
weren15	0.6932	0.8409	0.7727	0.2034	0.2000	0.2489	0.2003	0.1849	0.1831

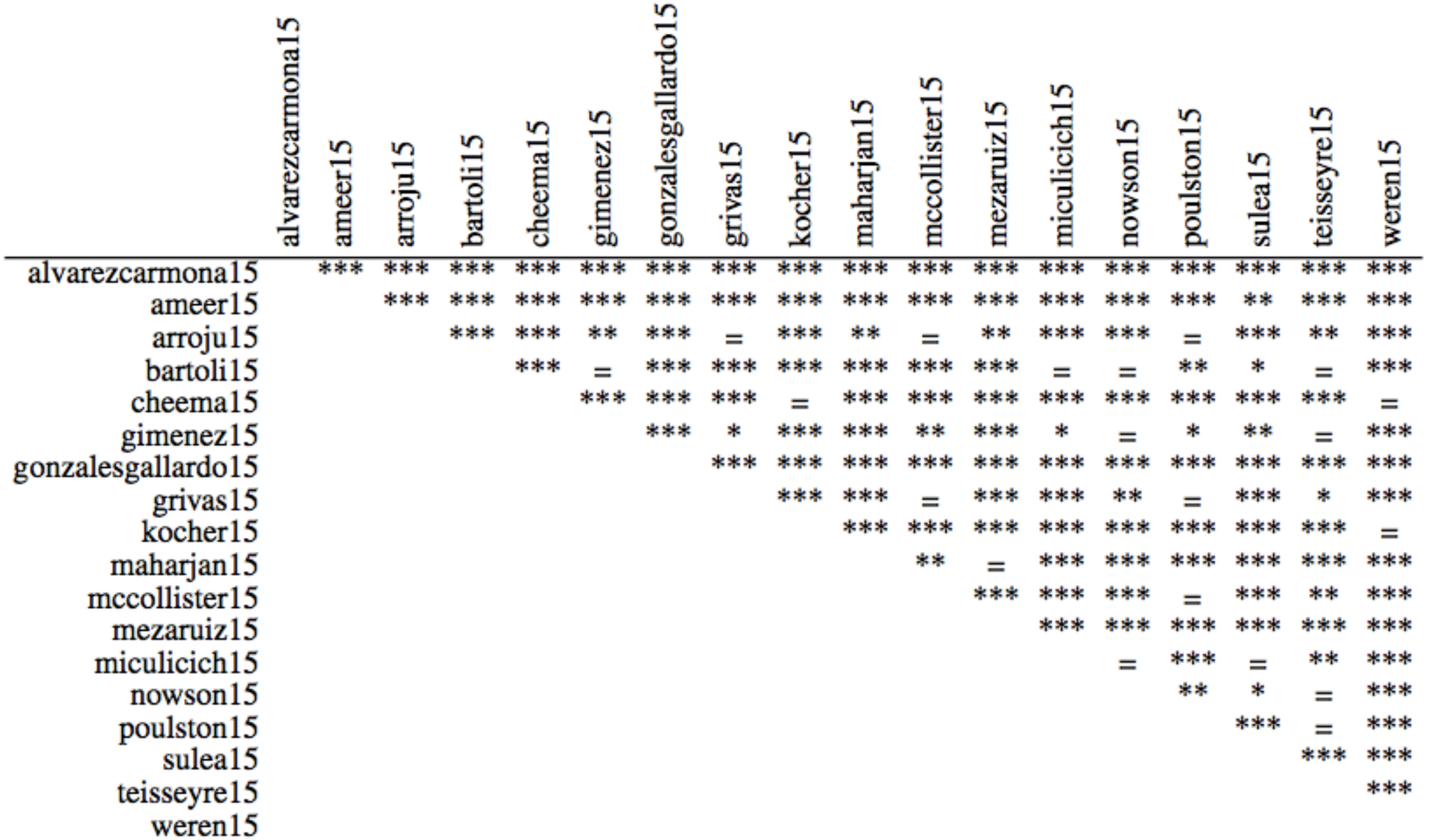
Italian

Team	Gender	*RMSE*	E	S	A	C	O
alvarezcarmona15	0.7222	**0.1044**	**0.0726**	0.1803	**0.0527**	0.1190	**0.0972**
ameer15	0.5833	0.1723	0.1067	0.2303	0.1462	0.1333	0.2449
arroju15	0.5833	0.1581	0.1480	0.1941	0.1520	0.1345	0.1620
bartoli15	0.5000	0.1405	0.1004	0.1889	0.1386	0.1298	0.1450
bayot15	0.5278	0.1989	0.1928	0.2349	0.1820	0.2173	0.1676
cheema15	0.5278	0.1730	0.1607	0.2205	0.1572	0.1364	0.1900
gimenez15	0.6944	0.1764	0.1394	0.2533	0.1624	0.1247	0.2021
gonzalesgallardo15	**0.8611**	0.1294	0.0764	0.2121	0.0745	0.1269	0.1572
grivas15	0.8333	0.1743	0.1350	0.1930	0.1389	0.2461	0.1586
kocher15	0.7778	0.1259	0.1000	**0.1555**	0.1302	**0.1093**	0.1344
maharjan15	0.6944	0.2122	0.1610	0.2181	0.2118	0.2225	0.2476
mccollister15	0.5556	0.1526	0.1296	0.1993	0.1471	0.1263	0.1610
mezaruiz15	0.5000	0.1636	0.1336	0.1997	0.1463	0.1553	0.1831
miculicich15	0.6389	0.1506	0.1093	0.1650	0.1202	0.1683	0.1900
nowson15	0.8056	0.1515	0.0905	0.2147	0.1237	0.1598	0.1686
poulston15	0.7500	0.1378	0.1279	0.1923	0.1257	0.1187	0.1243
sulea15	0.6389	0.1370	0.1141	0.1913	0.1220	0.1140	0.1438
teisseyre15	0.4167	0.2119	0.1616	0.2646	0.2173	0.1764	0.2398
weren15	0.5833	0.1732	0.1143	0.2593	0.1394	0.1344	0.2186

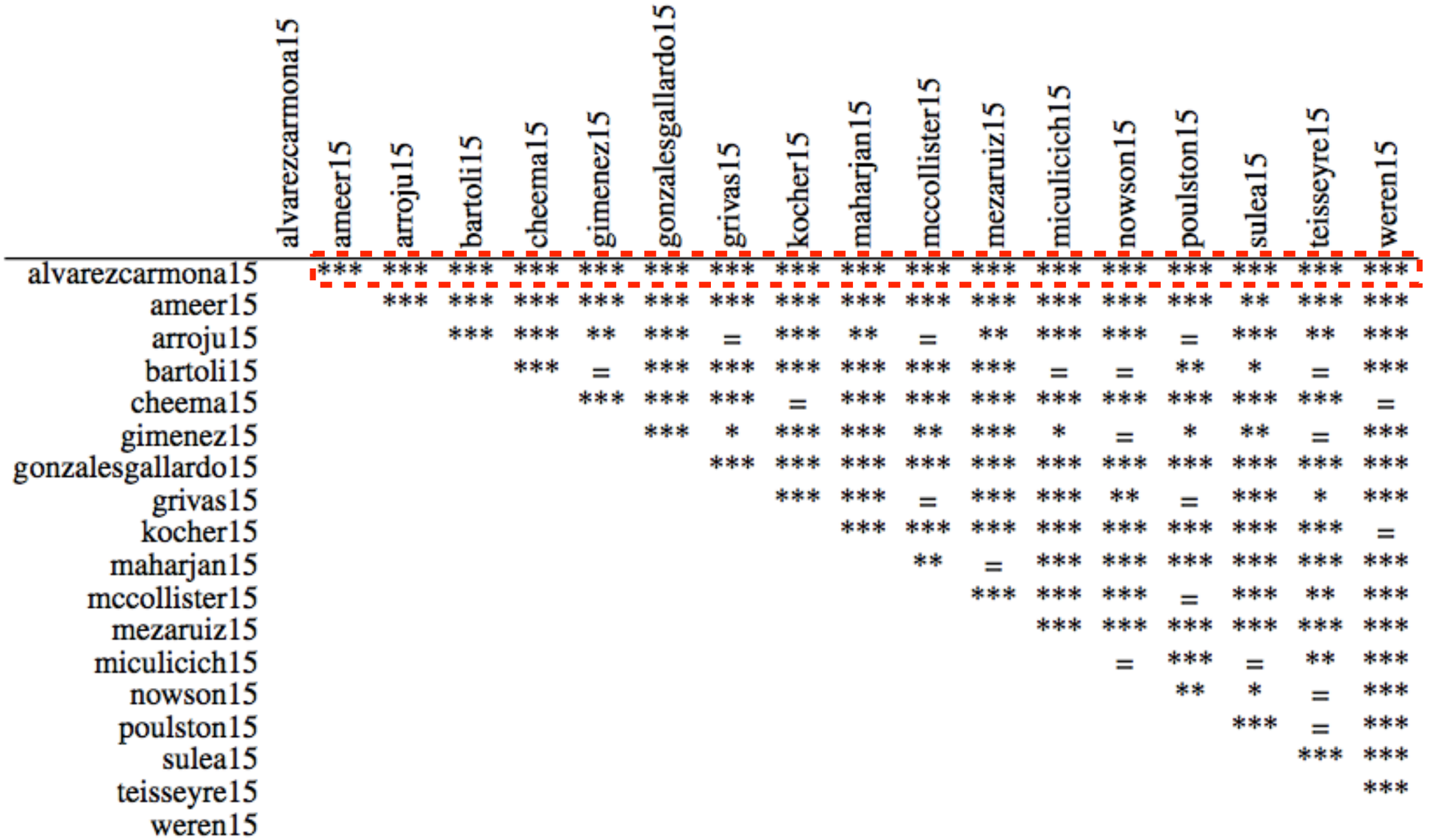
Dutch

Team	Gender	*RMSE*	E	S	A	C	O
alvarezcarmona15	0.9375	**0.0563**	**0.0750**	**0.0637**	**0.0000**	0.1075	**0.0354**
ameer15	0.5938	0.1418	0.1677	0.1686	0.1436	0.1425	0.0866
arroju15	0.5313	0.1627	0.1573	0.2235	0.1672	0.1553	0.1103
bartoli15	0.7188	0.1156	0.1467	0.1393	0.1261	0.0962	0.0696
bayot15	0.5625	0.1863	0.1705	0.2031	0.1631	0.1978	0.1969
cheema15	0.4688	0.1242	0.1369	0.1768	0.0919	0.1237	0.0919
gimenez15	0.7188	0.1607	0.1829	0.1785	0.1705	0.1392	0.1323
gonzalesgallardo15	0.9375	0.0890	0.0901	0.0661	0.0952	0.1299	0.0637
grivas15	**0.9688**	0.1571	0.1467	0.1711	0.1427	0.2278	0.0973
kocher15	0.8125	0.1186	0.1346	0.1225	0.1311	0.1299	0.0750
maharjan15	0.7813	0.2488	0.2102	0.2821	0.2781	0.2378	0.2358
markov15	0.5313	0.1716	0.1768	0.2411	0.1714	0.1436	0.1250
mccollister15	0.8125	0.1419	0.1499	0.1745	0.1497	0.1442	0.0913
mezaruiz15	0.5000	0.1595	0.1604	0.1928	0.1598	0.1787	0.1055
miculicich15	0.8125	0.1175	0.1199	0.1287	0.1046	0.1358	0.0984
nowson15	0.7813	0.1015	0.1350	0.1315	0.1086	**0.0619**	0.0703
poulston15	0.5000	0.1409	0.1752	0.1511	0.1444	0.1344	0.0993
sulea15	0.8438	0.1164	0.1310	0.1405	0.1114	0.1147	0.0846
teisseyre15	0.5938	0.1853	0.1862	0.2107	0.2187	0.1630	0.1479
weren15	0.6563	0.1491	0.1521	0.1620	0.1928	0.1323	0.1061

Pairwise comparison

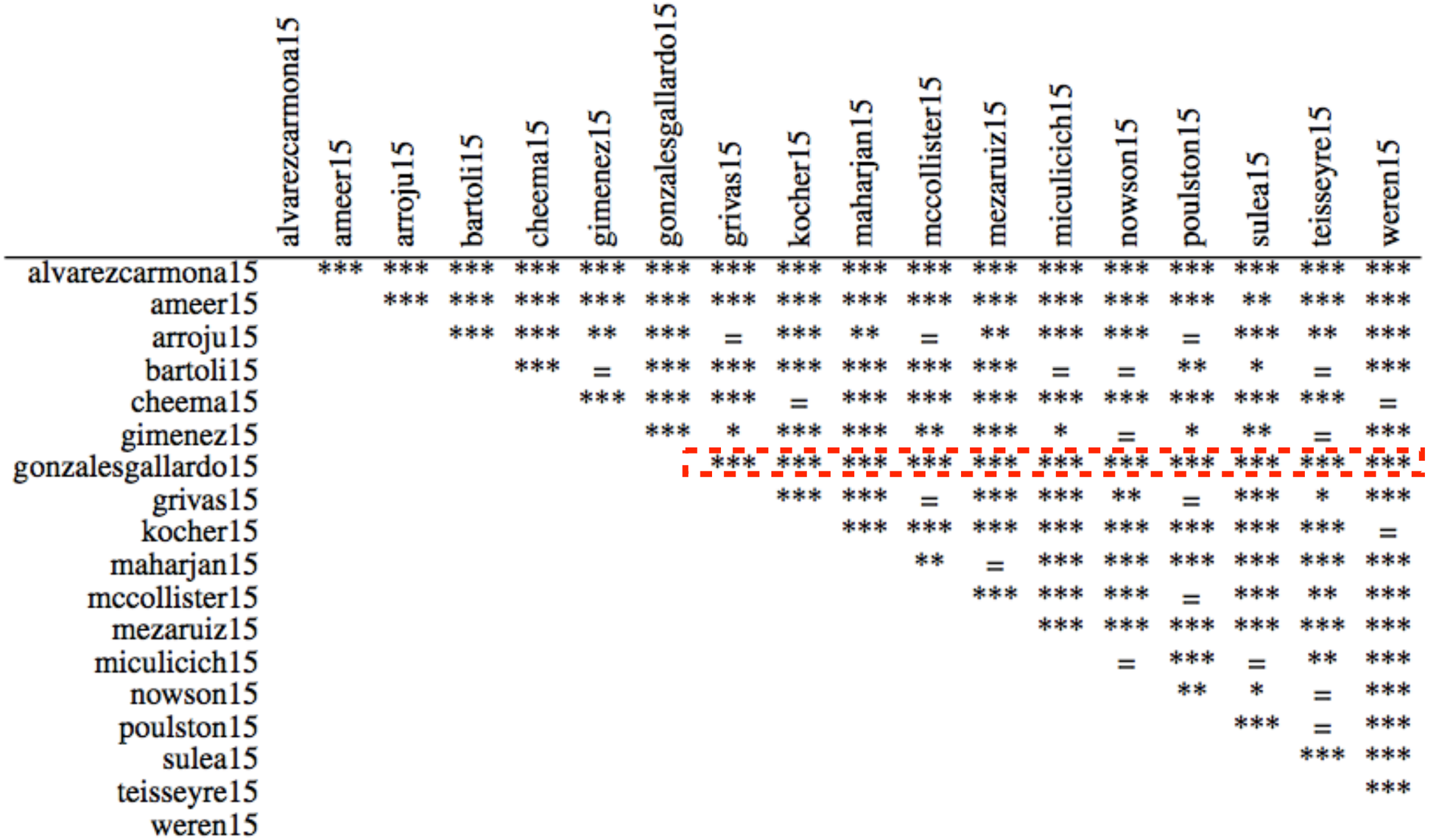


Pairwise comparison



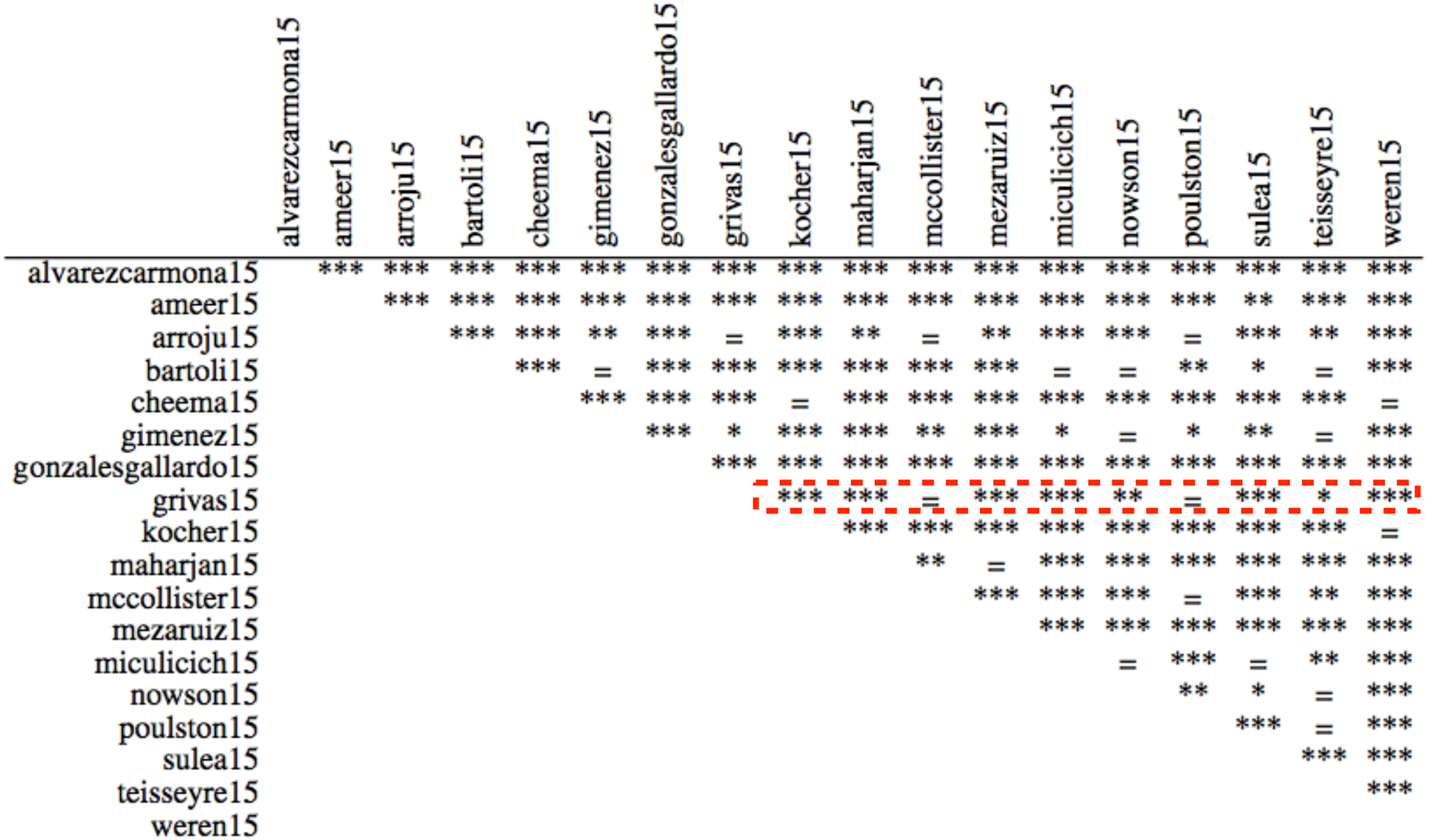
lest.

Pairwise comparison



2nd.

Pairwise comparison



3rd.

Conclusions

- Wrt. age and gender identification
  - The highest accuracies in gender identification were achieved in Dutch and Spanish with values over 95%
  - In comparison to previous years of PAN, the systems achieved significantly higher accuracy values for both age and gender identification.
  - This may suggest the number of tweets per author is sufficient to profile age and gender with high accuracy
- With regard to personality traits, the lowest errors were obtained for Dutch and Italian, with values below 5% for most traits.
  - The Stable trait appears the most difficult one to be predicted.
- Regarding the features it is difficult to highlight the most important ones, simply because the high number of different ones used and combined by the participants.
  - This year again the Second Order Representation proposed by alvarezcarmona15 obtained the best results.
  - representations based on n-grams, such as the one proposed by gonzalesgallardo15 or by grivas15, were ranked among the top three in every lan- guage.

Task impact

	CITATIONS
Overview of the Author Profiling task at PAN 2013	39
Overview of the 2nd. Author Profiling task at PAN 2014	16
Overview of the 3rd. Author Profiling task at PAN 2015	17

...and growing!!

Industry at PAN (Author Profiling)

Organisers	autoritas® nuevas ideas, nuevas soluciones
Sponsors	meaning cloud
Participants	xerox Xerox Research Centre Europe

Francisco Rangel

autoritas
nuevas ideas, nuevas soluciones

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Fabio Celli

UNIVERSITY OF TRENTO - Italy

Paolo Rosso

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Martin Potthast

Benno Stein

Bauhaus-Universität Weimar

Walter Daelemans

University of Antwerp

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