

Mining demographic info in social media: Age, gender and personality

Paolo Rosso



PRHLT Research Center
Universitat Politècnica de València

11th Int. Conf. on Intelligent Data Processing
Barcelona, 11th October 2016

Outline

- Author profiling: gender and age
- Author profiling in social media: shared tasks @ PAN



Author profiling

- Distinguishing between classes of authors, rather than individual authors
- **Marketing, Forensic Linguistics, Security**
 - Gender
 - Age
 - Personality profile: Big five personality traits
 - Native language
 - Language variety
 - Ideological/organizational affiliation
 - Etc.



Which is male/female?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

[examples: Moshe Koppel]

Distinguishing features: Male vs. female style

Males use more

- Determiners
- Adjectives
- *of* modifiers (e.g. *pot of gold*)

Informational
features

Females use more

- Pronouns
- *for* and *with*
- Negation
- Present tense

Involvedness
features

M. Koppel, S. Argamon, and A. R. Shimoni (2002). Automatically categorizing written texts by author gender. *Literary and linguistic computing* 17(4).



Which is male/female?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.



Which is male/female?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when **he** describes loose apposition as a rhetorical device. However, **he** does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does **he** specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper **I** follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As **I** have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. **Their** re-constructions are then compared with the original Hemingway version.

Which is male/female?



My aim in this article is to **show** that given a relevance theoretic approach to utterance interpretation, it is possible to **develop** a better understanding of what some of these so-called apposition markers **indicate**. It will be argued that the decision to **put** something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he **describes** loose apposition as a rhetorical device. However, he does not **justify** this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he **specify** what kind of effects might be achieved by a reformulation or explain how it **achieves** those effects. In this paper I **follow** Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker **recognises** that the original formulation did not **achieve** optimal relevance .

The main aim of this article is to **propose** an exercise in stylistic analysis which can be employed in the teaching of English language. It **details** the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Which is male/female?



My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does **not** justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. **Nor** does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did **not** achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are **not** as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.



Which is male/female?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Teen

Twenties

Thirties

Male

Female

Example

Yesterday we had our second jazz competition. Thank God we weren't competing. We were sooo bad. Like, I was so ashamed, I didn't even want to talk to anyone after. I felt so rotten, and I wanted to cry, but...it's ok.

Teen

Twenties

Thirties

Male

Female

Example

Yesterday we had our second jazz competition. Thank God we weren't competing. We were sooo bad. Like, I was so ashamed, I didn't even want to talk to anyone after. I felt so rotten, and I wanted to cry, but...it's ok.

Blog corpus

	Gender		
Age	Female	Male	Total
unknown	12287	12259	24546
13-17	6949	4120	8240
18-22	7393	7690	15083
23-27	4043	6062	8086
28-32	1686	3057	4743
33-37	860	1827	1720
38-42	374	819	748
43-48	263	584	526
>48	314	906	1220
Total	9660	9660	19320

Final balanced corpus:

- 19,320 total blogs
 - 8240 in “10s”
 - 8086 in “20s”
 - 2994 in “30s”
- 681,288 total posts
- 141,106,859 total words

J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker (2006). Effects of age and gender on blogging. In AACL Spring Symposium: Computational Approaches to Analyzing Weblogs, pages 199–205. AACL.

The lifecycle of the common blogger...

Word	10s	20s	30s
maths	105	3	2
homework	137	18	15
bored	384	111	47
sis	74	26	10
boring	369	102	63
awesome	292	128	57
mum	125	41	23
crappy	46	28	11
mad	216	80	53
dumb	89	45	22

The lifecycle of the common blogger...

Word	10s	20s	30s
maths	105	3	2
homework	137	18	15
bored	384	111	47
sis	74	26	10
boring	369	102	63
awesome	292	128	57
mum	125	41	23
crappy	46	28	11
mad	216	80	53
dumb	89	45	22

Word	10s	20s	30s
semester	22	44	18
apartment	18	123	55
drunk	77	88	41
beer	32	115	70
student	65	98	61
album	64	84	56
college	151	192	131
someday	35	40	28
dating	31	52	37
bar	45	153	111

The lifecycle of the common blogger...

Word	10s	20s	30s
maths	105	3	2
homework	137	18	15
bored	384	111	47
sis	74	26	10
boring	369	102	63
awesome	292	128	57
mum	125	41	23
crappy	46	28	11
mad	216	80	53
dumb	89	45	22

Word	10s	20s	30s
semester	22	44	18
apartment	18	123	55
drunk	77	88	41
beer	32	115	70
student	65	98	61
album	64	84	56
college	151	192	131
someday	35	40	28
dating	31	52	37
bar	45	153	111

Word	10s	20s	30s
marriage	27	83	141
development	16	50	82
campaign	14	38	70
tax	14	38	72
local	38	118	185
democratic	13	29	59
son	51	92	237
systems	12	36	55
provide	15	54	69
workers	10	35	46

Men are from Mars... Women are from Venus...

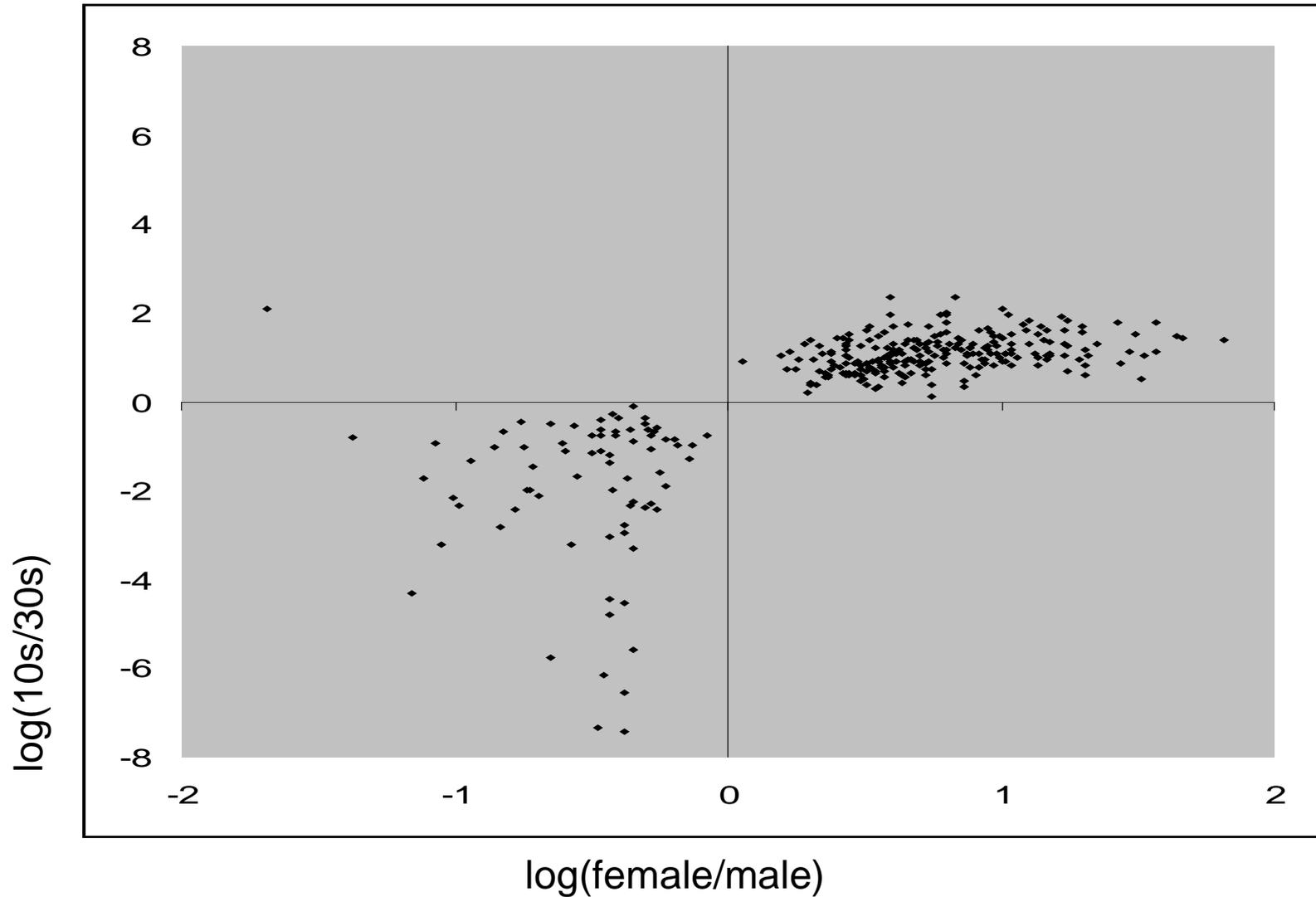


LIWC category	male	female
job	<u>68.1±0.6</u>	56.5±0.5
money	<u>43.6±0.4</u>	37.1±0.4
sports	<u>31.2±0.4</u>	20.4±0.2
tv	<u>21.1±0.3</u>	15.9±0.2
sex	32.4±0.4	<u>43.2±0.5</u>
family	27.5±0.3	<u>40.6±0.4</u>
eating	23.9±0.3	<u>30.4±0.3</u>
friends	20.5±0.2	<u>25.9±0.3</u>
sleep	18.4±0.2	<u>23.5±0.2</u>
<i>pos-emotions</i>	248.2±1.9	<u>265.1±2</u>
<i>neg-emotions</i>	159.5±1.3	<u>178±1.4</u>

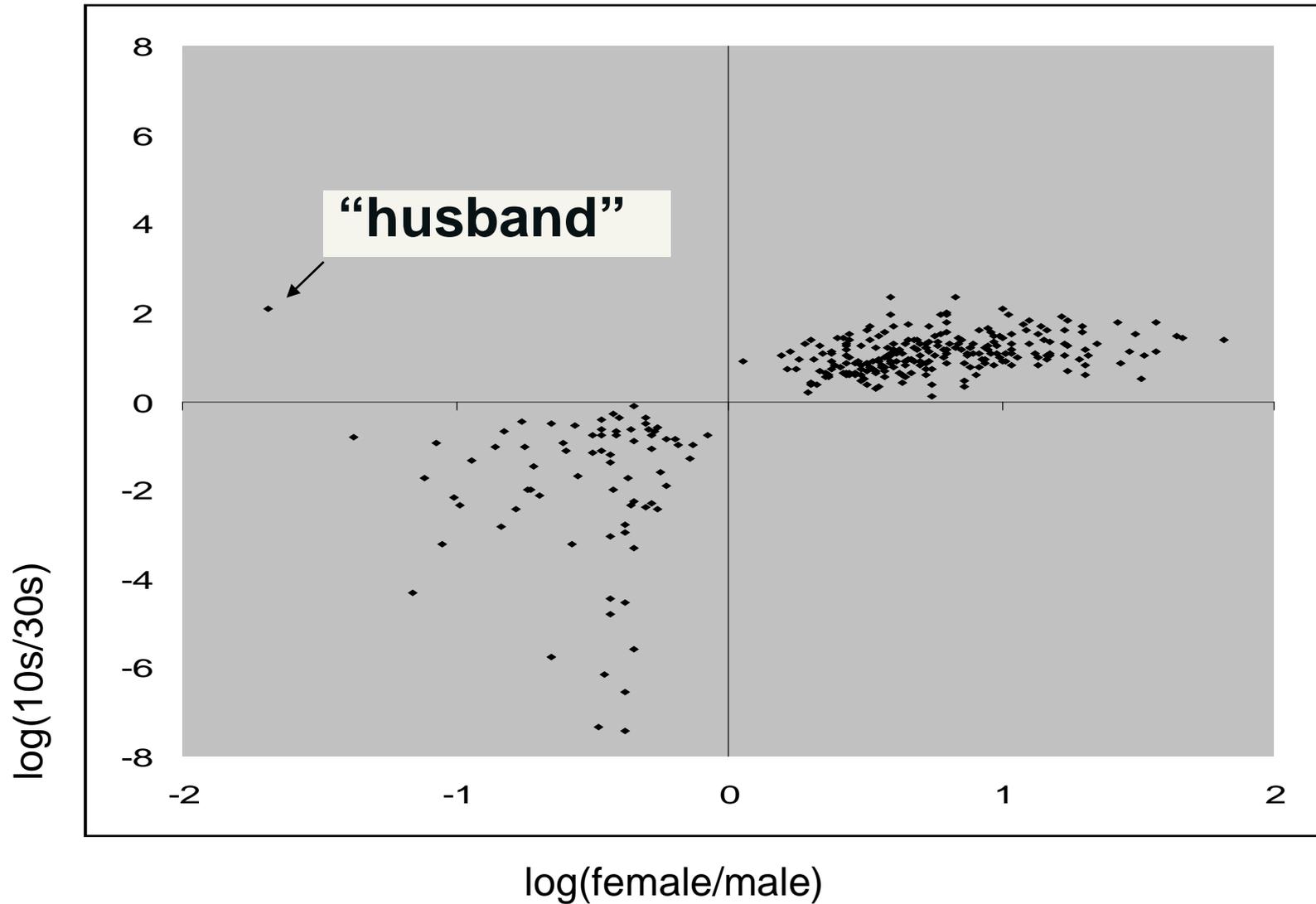
Relating age & gender

- Now...is there a linguistic connection between age and gender?
- Consider the most distinctive words for both Age and Gender:
 - Intersect the 1000 words with **highest Age information gain** and the 1000 words with **highest Gender information gain**
 - Total of 316 words
 - Plot $\log(30s/10s)$ vs. $\log(\text{male/female})$

Relating age & gender



Relating age & gender





PAN

Uncovering plagiarism,

authorship, and social software misuse

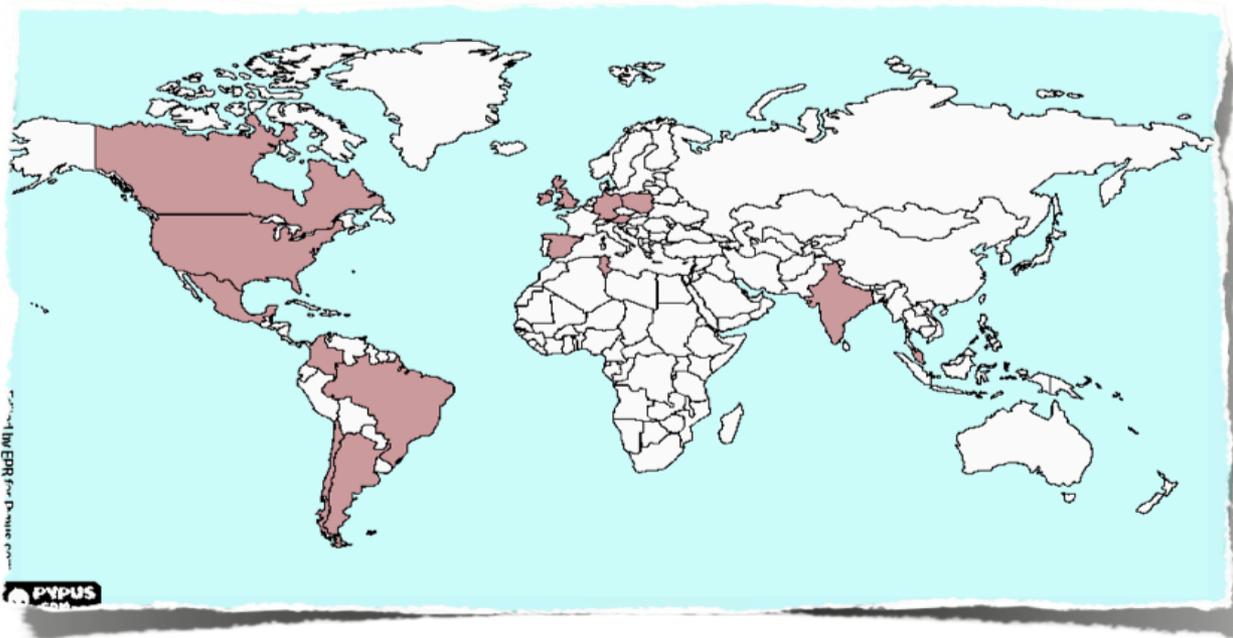
Since 2007 as workshop (SIGIR, ECAI); since 2009 as @ Conference and Labs of the Evaluation Forum (**CLEF**) and Forum of the

Information Retrieval Evaluation (**FIRE**): <http://pan.webis.de/>

- Plagiarism detection (since 2009)
- Author identification (since 2011)
- **Author profiling** (since 2013)
- Online sexual predator (2012)
- Author obfuscation (2016)

PAN Author profiling @ CLEF-13

- Teams submitting results: 21 (registered teams: 64)
- (Towards) **big data**: 400,000 social media texts including **chat lines of potential pedophiles** (task @ PAN-12)



- **Age classes**: 10s (13-17), 20s (23-27), 30s (33-48)
- **Languages**: English and Spanish

English			
Team	Total	Gender	Age
Meina	0.3894	0.5921	0.6491
Pastor L.	0.3813	0.5690	0.6572
Seifeddine	0.3677	0.5816	0.5897
Santosh	0.3508	0.5652	0.6408
Yong Lim	0.3488	0.5671	0.6098
Ladra	0.3420	0.5608	0.6118
Aleman	0.3292	0.5522	0.5923
Gillam	0.3268	0.5410	0.6031
Kern	0.3115	0.5267	0.5690
Cruz	0.3114	0.5456	0.5966
Pavan	0.2843	0.5000	0.6055
Caurcel Diaz	0.2840	0.5000	0.5679
H. Farias	0.2816	0.5671	0.5061
Jankowska	0.2814	0.5381	0.4738
Flekova	0.2785	0.5343	0.5287
Weren	0.2564	0.5044	0.5099
Sapkota	0.2471	0.4781	0.5415
De-Arteaga	0.2450	0.4998	0.4885
Moreau	0.2395	0.4941	0.4824
baseline	0.1650	0.5000	0.3333
Gopal Patra	0.1574	0.5683	0.2895
Cagnina	0.0741	0.5040	0.1234

Spanish			
Team	Total	Gender	Age
Santosh	0.4208	0.6473	0.6430
Pastor L.	0.4158	0.6299	0.6558
Cruz	0.3897	0.6165	0.6219
Flekova	0.3683	0.6103	0.5966
Ladra	0.3523	0.6138	0.5727
De-Arteaga	0.3145	0.5627	0.5429
Kern	0.3134	0.5706	0.5375
Yong Lim	0.3120	0.5468	0.5705
Sapkota	0.2934	0.5116	0.5651
Pavan	0.2824	0.5000	0.5643
Jankowska	0.2592	0.5846	0.4276
Meina	0.2549	0.5287	0.4930
Gillam	0.2543	0.4784	0.5377
Moreau	0.2539	0.4967	0.5049
Weren	0.2463	0.5362	0.4615
Cagnina	0.2339	0.5516	0.4148
Caurcel Diaz	0.2000	0.5000	0.4000
H. Farias	0.1757	0.4982	0.3554
baseline	0.1650	0.5000	0.3333
Aleman	0.1638	0.5526	0.2915
Seifeddine	0.0287	0.5455	0.0512
Gopal Patra	–	–	–

Approaches: Features

- Stylistic: frequency of punctuation marks, capital letters,...
- Part of Speech
- Readability measures
- Dictionary-based words, topic-based words
- Collocations
- Character or word n-grams
- Slang words, character flooding
- Emoticons
- Emotion words

F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches (2013). Overview of the Author Profiling Task at PAN 2013 - Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings Vol. 1179.

Graph-based features

Given a graph $G=\{N,E\}$ where:

- N is the set of nodes
- E is the set of edges

we obtain a set of:

- structure-based features from global measures of the graph
- node-based features from node specific measures

Structure-based features

Nodes-edges ratio	Indicator of how connected the graph is, i.e., how complicated the discourse is.	Theoretical maximum: $\max(E) = N * (N - 1)$
Weighted average degree	Indicator of how much interconnected the graph is, i.e., how much interconnected the grammatical categories are.	Averaging all nodes degrees. Scaling it to [0,1]
Diameter	Indicator of the greatest distance between any pair of nodes, i.e, how far a grammatical category is from others , or how far a topic is from an emotion.	$d = \max_{n \in N} \epsilon(N)$ where $E(N)$ is the eccentricity
Density	Indicator of how close the graph is to be complete, i.e., how dense is the text in the sense of how each grammatical category is used in combination to others.	$D = \frac{2 * E }{(N * (N - 1))}$
Modularity	Indicator of different divisions of the graph into modules (one node has dense connections within the module and sparse with nodes in other modules), i.e., how the discourse is modeled in different structural or stylistic units.	Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E. Fast unfolding of communities in large networks. In: Journal of Statistical Mechanics: Theory and Experiment, vol. 2008 (10), pp. 10008 (2008)
Clustering coefficient	Indicator of the transitivity of the graph (if a is directly linked to b and b is directly linked to c, what's the probability that a node is directly linked to c), i.e., how different grammatical categories or semantic information are related to each other	Watts-Strogatz: $cc1 = \frac{\sum_{i=1}^n C(i)}{n}$
Average path length	Indicator of how far some nodes are from others, i.e., how far some grammatical categories are from others, or for example how far some topics are from some emotions	Brandes, U. A Faster Algorithm for Betweenness Centrality. In: Journal of Mathematical Sociology 25(2), pp. 163-177 (2001)

Node-based features

<p>EigenVector</p>	<p>It gives a measure of the influence of each node. In our case, it may give what are the grammatical categories with the most central use in the author's discourse, for example, which nouns, verbs or adjectives</p>	<p>Given a graph and its adjacency matrix $A = a_{n,t}$ where $a_{n,t}$ is 1 if a node n is linked to a node t, and 0 otherwise:</p> $x_n = \frac{1}{\lambda} \sum_{t \in M(n)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{n,t} x_t$ <p>where λ is a constant representing the greatest eigenvalue associated with the centrality measure.</p>
<p>Betweenness</p>	<p>It gives a measure of the importance of a each node depending on the number of shortest paths of which it is part of. In our case, if one node has a high betweenness centrality means that it is a common element used for link among parts-of-speech, for example, prepositions, conjunctions or even verbs and nouns. Hence, this measure may give us an indicator of how the most common connectors in the linguistic structures used by authors</p>	<p>It is the ratio of all shortest paths from one node to another node in the graph that pass through x:</p> $BC(x) = \sum_{i,j \in N - \{n\}} \frac{\sigma_{i,j}(n)}{\sigma_{i,j}}$ <p>Where $\sigma_{i,j}$ is the total number of shortest paths from node i to j, and $\sigma_{i,j}(n)$ is the total number of those paths that pass through n.</p>

EmoGraph @ CLEF-13

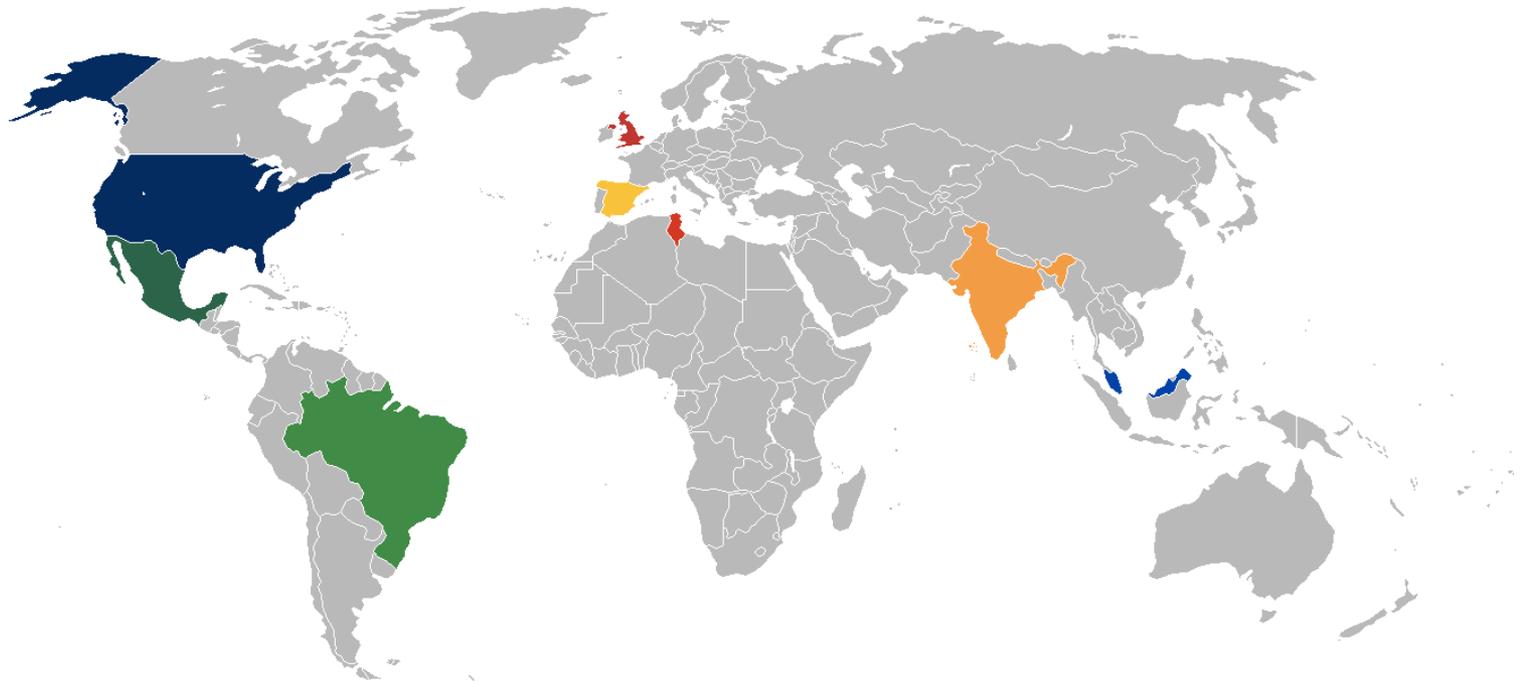
Ranking	Team	Accuracy
1	Rangel-EG	0.6624
2	Pastor	0.6558
3	Santosh	0.6430
4	Rangel-S	0.6350
5	Haro	0.6219
6	Flekova	0.5966
...	...	
21	Baseline	0.3333
...	...	
23	Mechti	0.0512

Ranking	Team	Accuracy
1	Santosh	0.6473
2	Rangel-EG	0.6365
3	Pastor	0.6299
4	Haro	0.6165
5	Ladra	0.6138
...	...	
8	Rangel-S	0.5713
...	...	
18	Baseline	0.5000
...	...	
23	Gillam	0.4784

F. Rangel, P. Rosso (2016). On the impact of emotions on author profiling. Information Processing & Management, 52(1): 73-92.

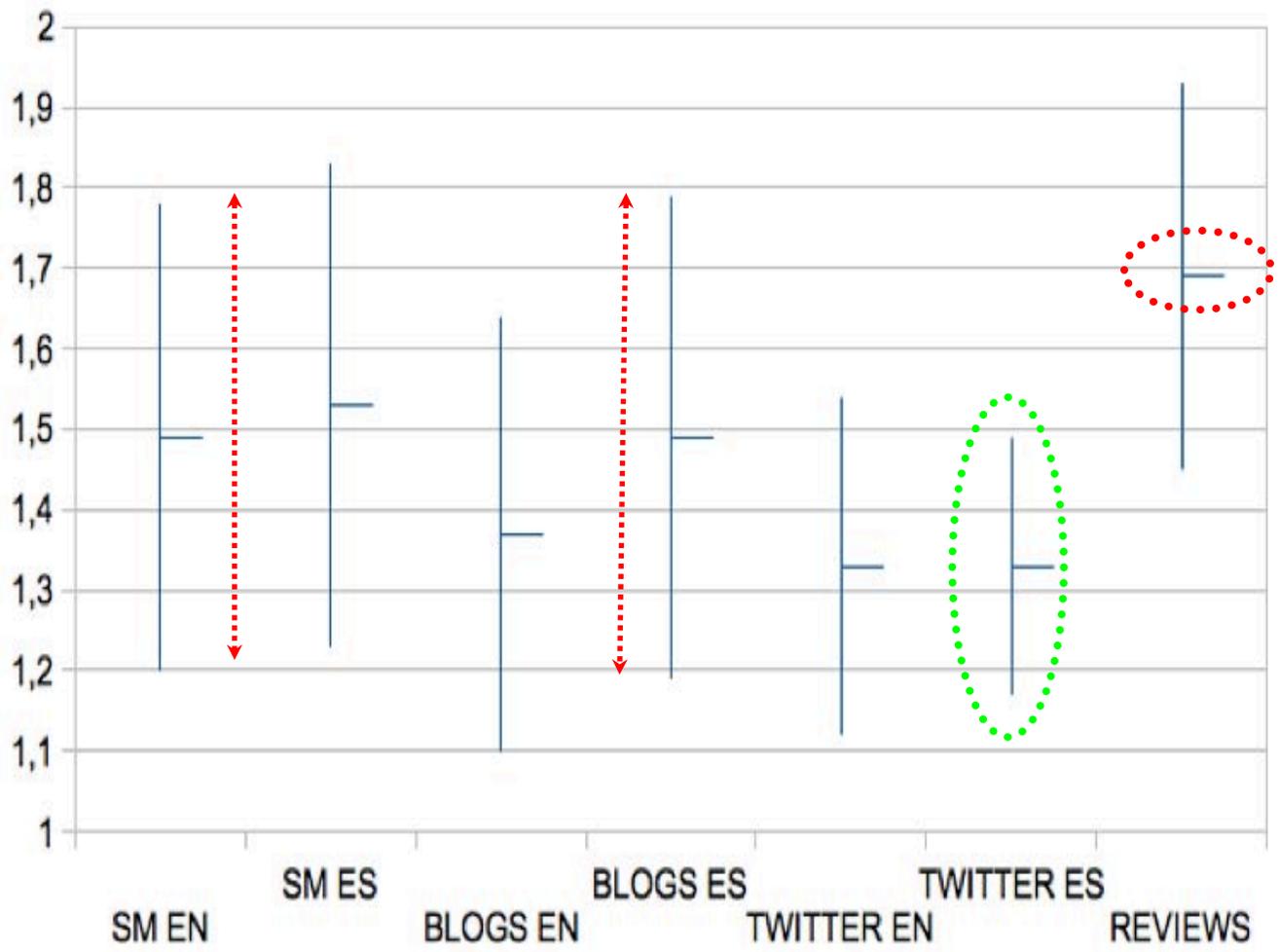
PAN Author profiling @ CLEF-14

- Teams submitting results: 10
- **Social media + blogs + Twitter + TripAdvisor**
- **Age classes: 18-24, 25-34, 35-49, 50-64, 65+**
- Languages: English and Spanish





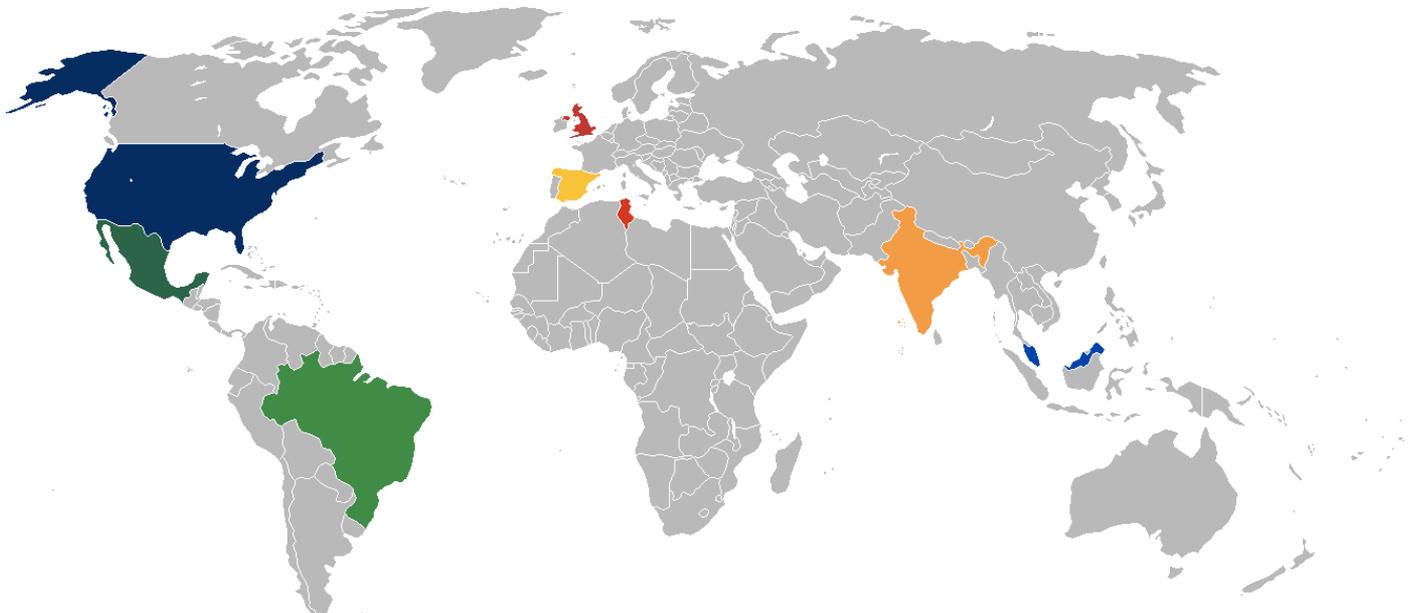
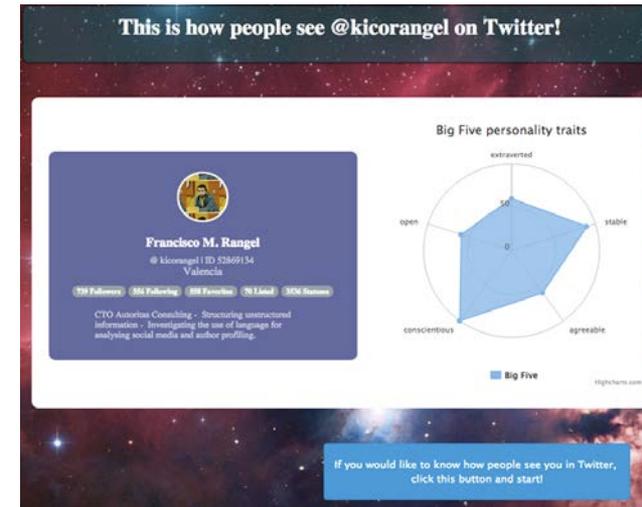
F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkman, B. Stein, B. Verhoeven, and W. Daelemans (2014). Overview of the 2nd Author Profiling Task at PAN 2014—Notebook for PAN at CLEF 2014. CEUR Workshop Proceedings Vol. 1180, pp. 898-927.



Twitter: more spontaneous way to communicate

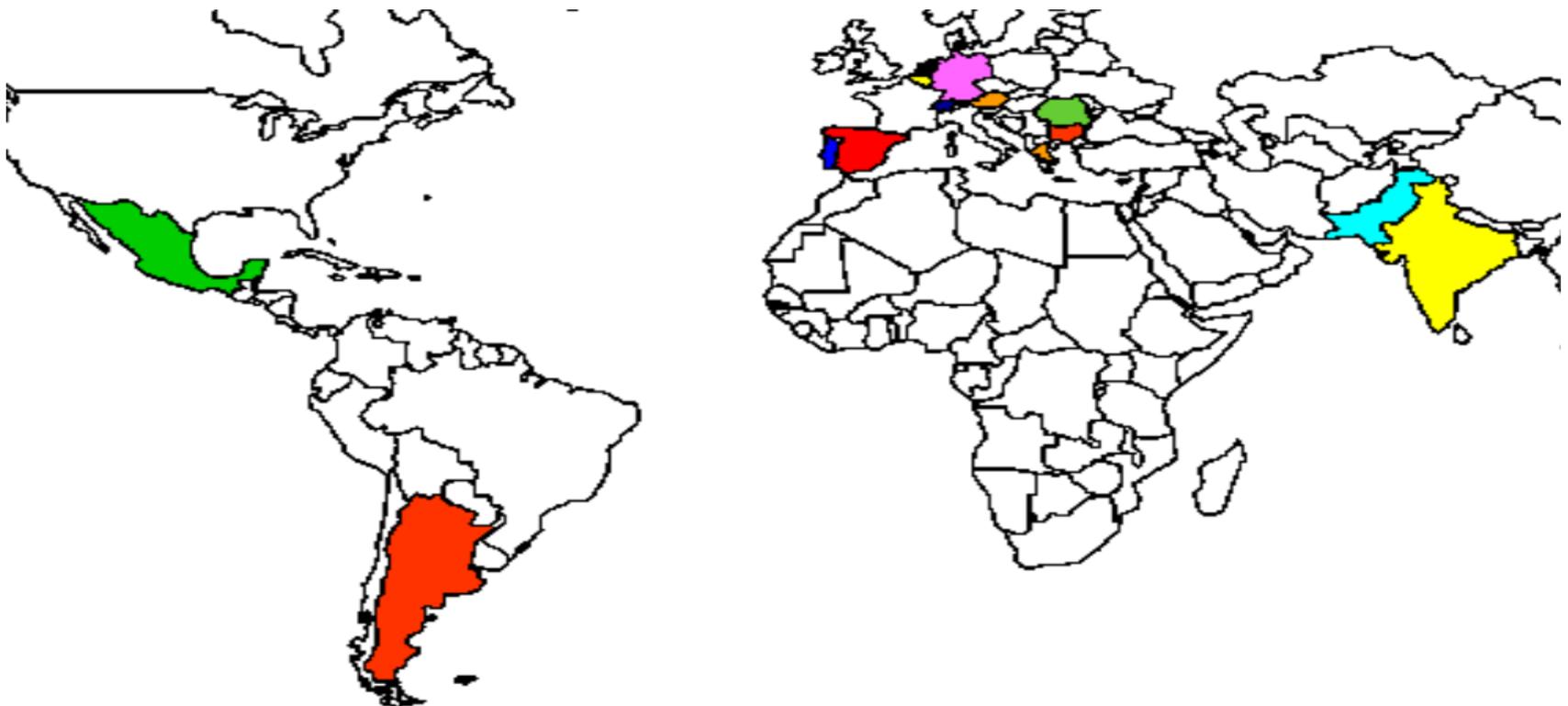
PAN Author profiling @ CLEF-15

- Teams submitting results: 22
- Age classes: 18-24, 25-34, 35-49, 50+
- Gender, age and **personality** in **Twitter**
- <http://your-personality-test.com/>
- Languages: English, Spanish, Italian, Dutch



PAN Author profiling @ CLEF-16

- Teams submitting results: 22
- **Cross-genre** gender and age (train: **Twitter**; test: **blogs**, etc)
- Age classes: 18-24, 25-34, 35-49, 35-49, 50-64, 65+
- Languages: English, Spanish, Dutch





PAN Author profiling @ FIRE-16

- **PR-SOCO: Personality Recognition in Source COde**
- Teams submitting results: 11 (49 runs)
- Big-5: Open, Conscientious, Neurotic, Extroverted, Agreeable

Author profiling @ FIRE-17 ?

- **RusProfiling: Author profiling in Russian**
- *RusPersonality* corpus of Russian texts with metadata demographic info (**Tatiana Litvinova**, head of Corpus Sociolinguistics and Authorship Profiling Lab, Voronezh)
- *Organisers*: Tatiana Litvinova , Olga Zagorovskaya, Olga Litvinova, Pavel Seredin (Voronezh); Olga Romanchenko, Evgeny Larin...

Thanks / Gracias / Spasibo
Any question?

Paolo Rosso

IDP-2016

proso@dsic.upv.es

<http://www.dsic.upv.es/~proso/>

J. W. Pennebaker. (2013)

The secret life of pronouns: What our words say about us.
Bloomsbury USA.