Profiling information in social media

Paolo Rosso
PRHLT Research Center
Universitat Politècnica de València
http://www.dsic.upv.es/~prosso/

Tutorial @ CLiC-it
5th Italian Conference on Computational Linguistics
Torino 10/12/2018

Outline

- Profiling gender & age
- Author profiling shared tasks at PAN
- EmoGraph: graph-based discourse analysis
- Profiling native language
- Profiling sexual offenders
- Profiling irony

Author Profiling

Language and style varies among classes of authors

Forensics: who is behind an harassment

Security: who is behind a threat

Marketing: who is behind an opinion

Socio-political analysis: who is behind a stance

- **Gender & age**
- **Personality**
- **Native language and language variety**
- Ideological/organizational affiliation

Security


ARABIC AUTHOR PROFILING
FOR CYBER SECURITY
FUNDED BY THE QATAR NATIONAL RESEARCH FUND

Profiling for security

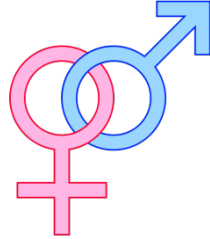
- Profiling a deceiver
- Profiling irony

In case of a potential threat:

- Profiling gender
- Profiling age
- Profiling native language
- Profiling language variety

Profiling gender & age

Which is female/male?



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]

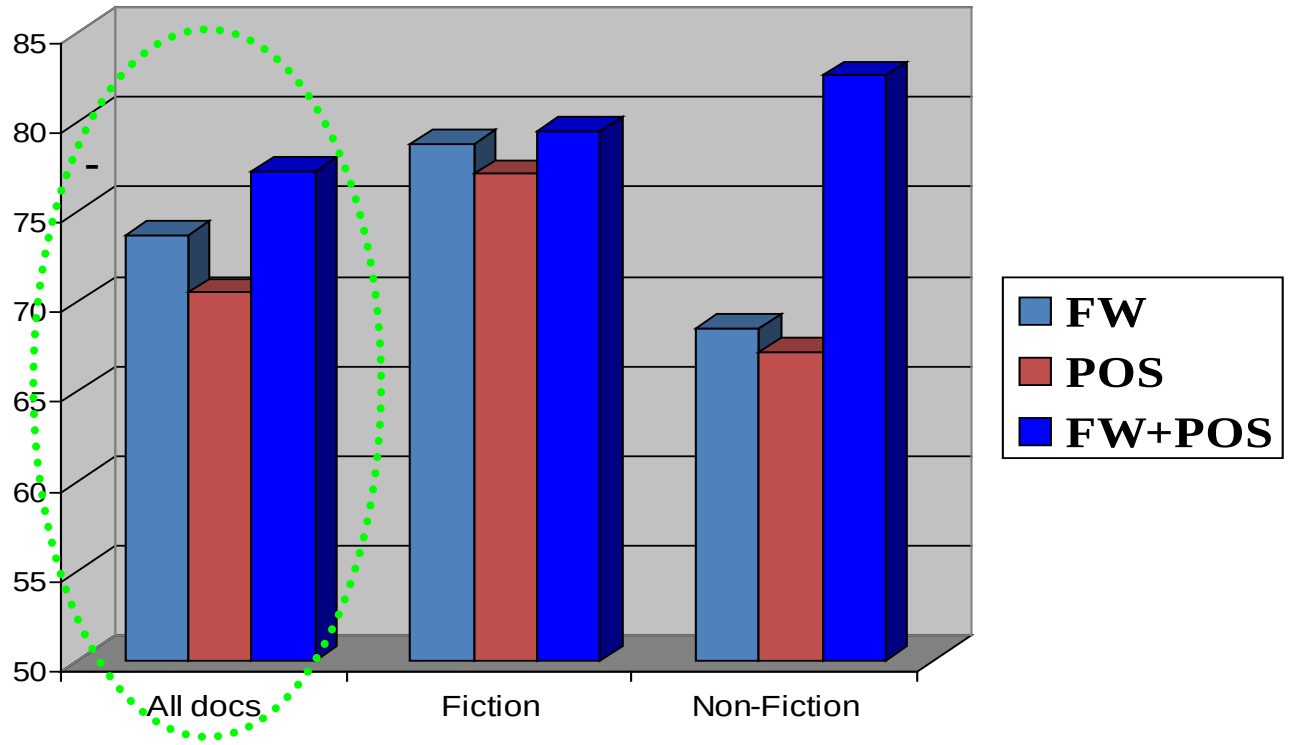
British National Corpus

- 920 documents labelled for
  - author gender
  - document genre
- Used 566 controlled for genre

	Male	Fem
Fiction (prose)	**132**	**132**
Non-fiction	**151**	**151**
Arts (general)	8	8
Arts (acad.)	12	12
Belief/Thought	12	12
Biography	27	27
Commerce	5	5
Leisure	8	8
Science (gen.)	13	13
Soc. Sci. (gen.)	26	26
Soc. Sci. (acad.)	19	19
World Affairs	21	21

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

Results per feature set



- Handle fiction and non-fiction separately
- Use full feature set

POS: Part Of Speech FW: Function words (*and, of, the*,..)

Distinguishing features: male vs. female style

Males use more

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

Informational features

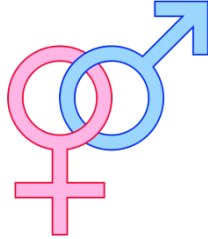
Females use more

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

Involvedness features

J. W. Pennebaker. The Secret Life of Pronouns: What Our Words Say about Us. Bloomsbury USA, 2013.

Which is female/male?



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.

Female vs. male

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.

Female vs. male

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.

Female vs. male

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.

Female vs. male

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

Social media: 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 rotton, and I wanted to cry, but...it's ok.

Teen Twenties Thirties

Male Female

Social media: 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 rotton, and I wanted to cry, but...it's ok.

Blog corpus

- Less-formal text
  - 85,000 blogs
  - blogger-provided profiles (gender, age, occupation, astrological sign)
  - harvested August 2004
  - all non-text ignored (formatting, quoting)

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

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”

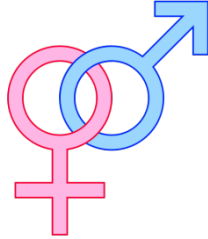
- 681,288 total posts
- 141,106,859 total words

Gender and age classification

Features	Gender & age (accuracy)
Style & Content	80.0% - 77.4%
Style Words	77.0% - 69.4%
Content Words	73.0% - 76.2%

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

Men vs. women



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

J. W. Pennebaker - LIWC: Linguistic Inquiry and Word Count

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

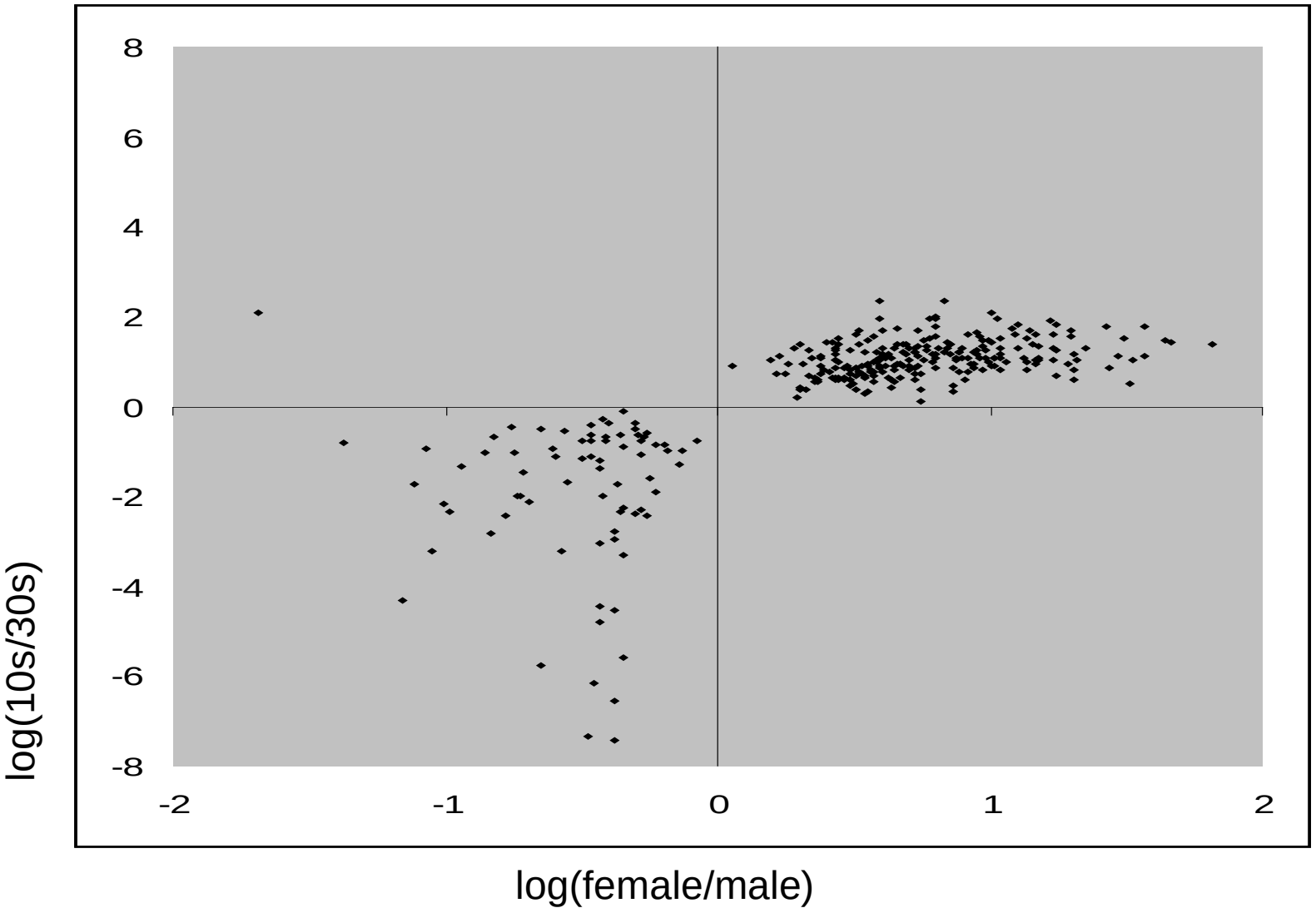
Word	10s	20s	30s
semester	22	44	18
apartment	18	123	55
drunk	77	88	41
beer	32	115	70
student	65	98	61
album	64	84	56
college	151	192	131
someday	35	40	28
dating	31	52	37
bar	45	153	111

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

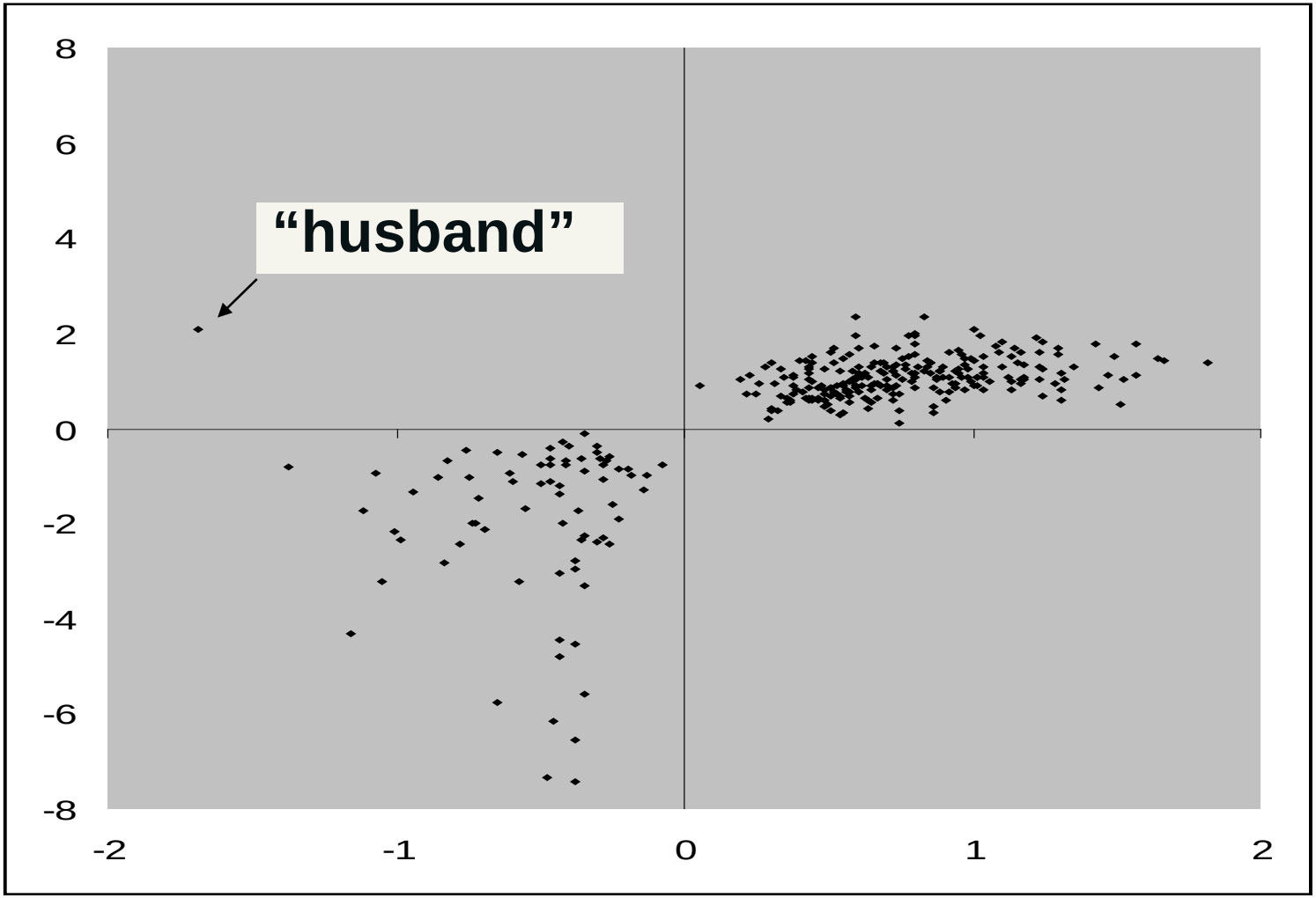
Relating age & gender

- 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(male/female)

Relating age & gender



Relating age & gender



Gender & age: pre PAN state of the art

AUTHOR	COLLECTION	FEATURES	RESULTS	OTHER CHARACTERISTICS
Argamon et al., 2002	British National Corpus	Part-of-speech	Gender: 80% accuracy	
Koppel et al., 2003	Blogs	Lexical and syntactic features	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

Author profiling @ PAN

Francisco Rangel, Autoritas & Universitat Politècnica de València
Moshe Koppel, Bar-Illan University
Efstathios Stamatatos, University of the Aegean
Walter Daelemans, University ofAntwerp
Fabio Celli, University of Trento
…
Paolo Rosso, Universitat Politècnica de València

PAN digital text forensics and sylometry

Since 2007 as workshop (SIGIR, ECAI)

Since 2009 organizing benchmark activities: http://pan.webis.de/

since 2010 @ Conference and Labs of the Evaluation Forum (CLEF)

since 2011 also @ Forum of Information Retrieval Evaluation (FIRE)

Plagiarism detection (since 2009)

Author identification (since 2011)

Author profiling (since 2013)

Online sexual predator (in 2012)

Author obfuscation (in 2016)

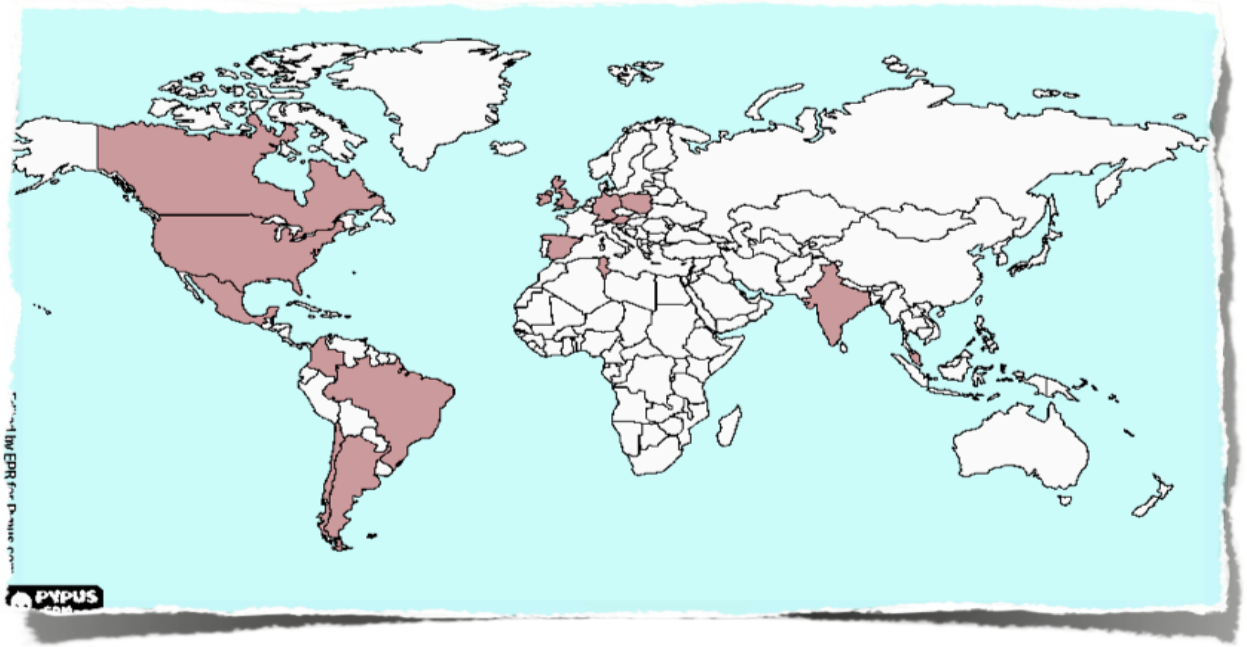
Author profiling @ PAN

- CLEF 2013: Age and gender in social media
- CLEF 2014: Age and gender in social media, Twitter, blogs, reviews
- CLEF 2015: Age, gender, personality in Twitter
- CLEF 2016: Cross-genre age and gender
- FIRE 2016: Personality in source code
- CLEF 2017: Gender and language variety identification in Twitter
- FIRE 2017: Native Indian language identification
- FIRE 2017: Cross-genre gender identification in Russian
- CLEF 2018: Multimodal (text + image) age and gender in Twitter

Author profiling: PAN @CLEF 2013

- Teams submitting results: 21 (Registered teams: 64)
- (Towards) **big data**: 400,000 social media texts

including **chat lines of potential pedophiles** (task in 2012)



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

Results: EN vs. ES

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	–	–	–

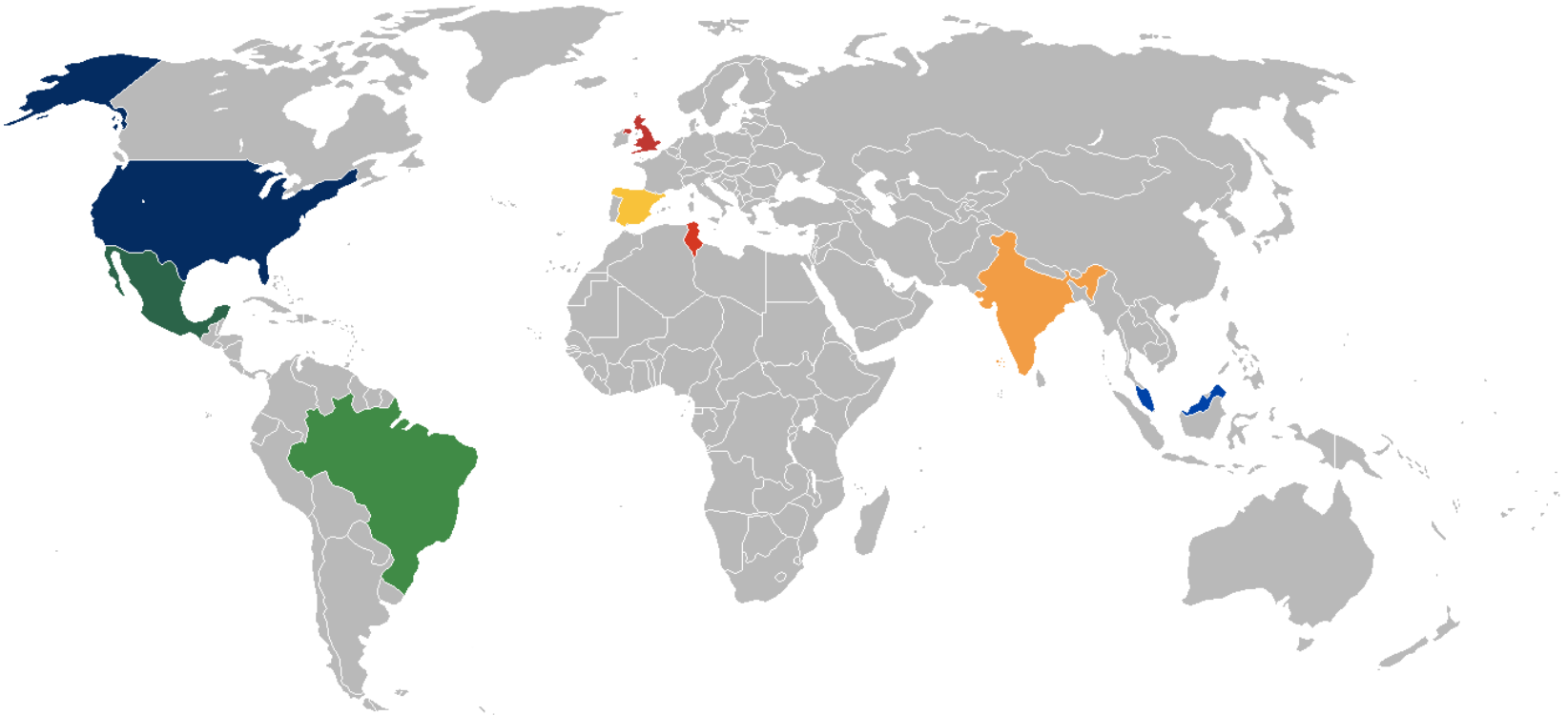
Features

- 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. Overview of the Author Profiling Task at PAN 2013 - Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings Vol. 1179. 2013.

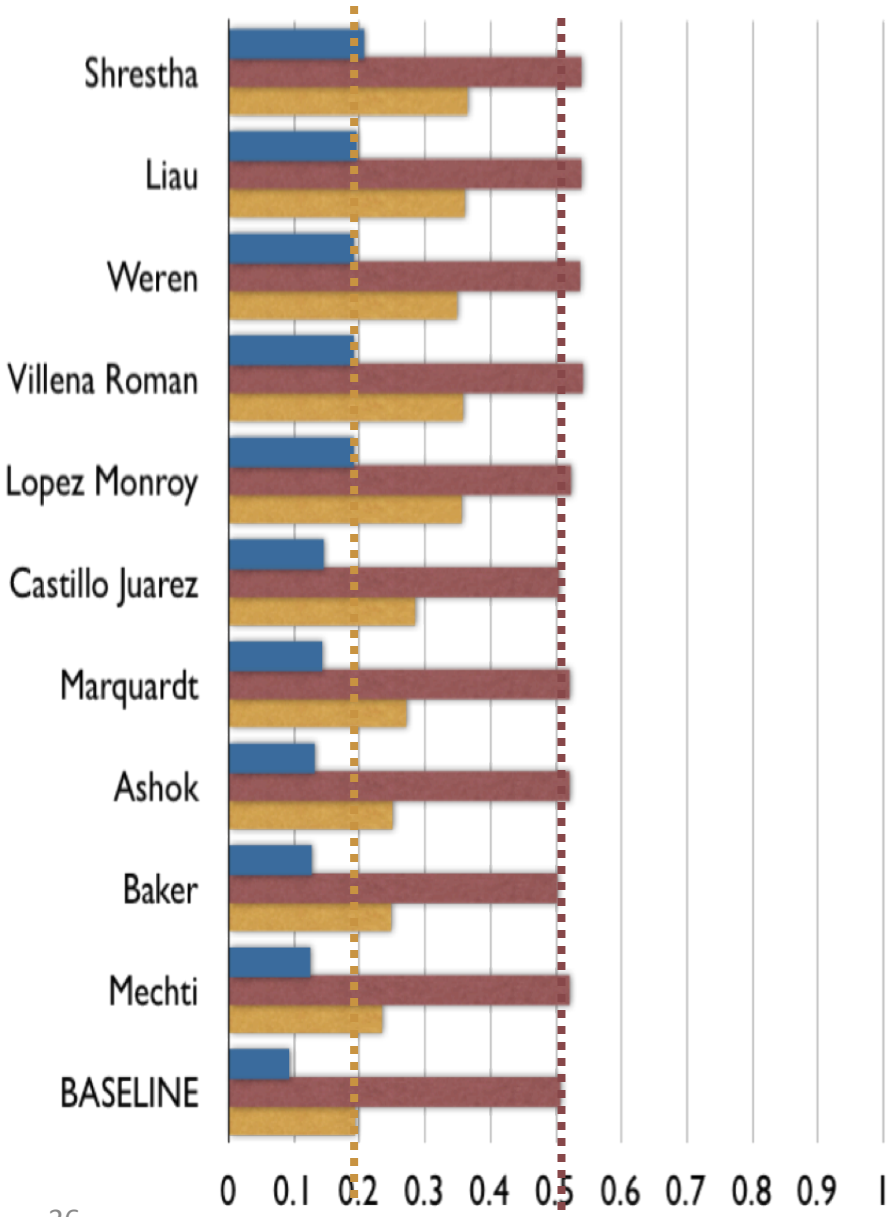
Author profiling PAN @CLEF 2014

- Teams submitting results: 10

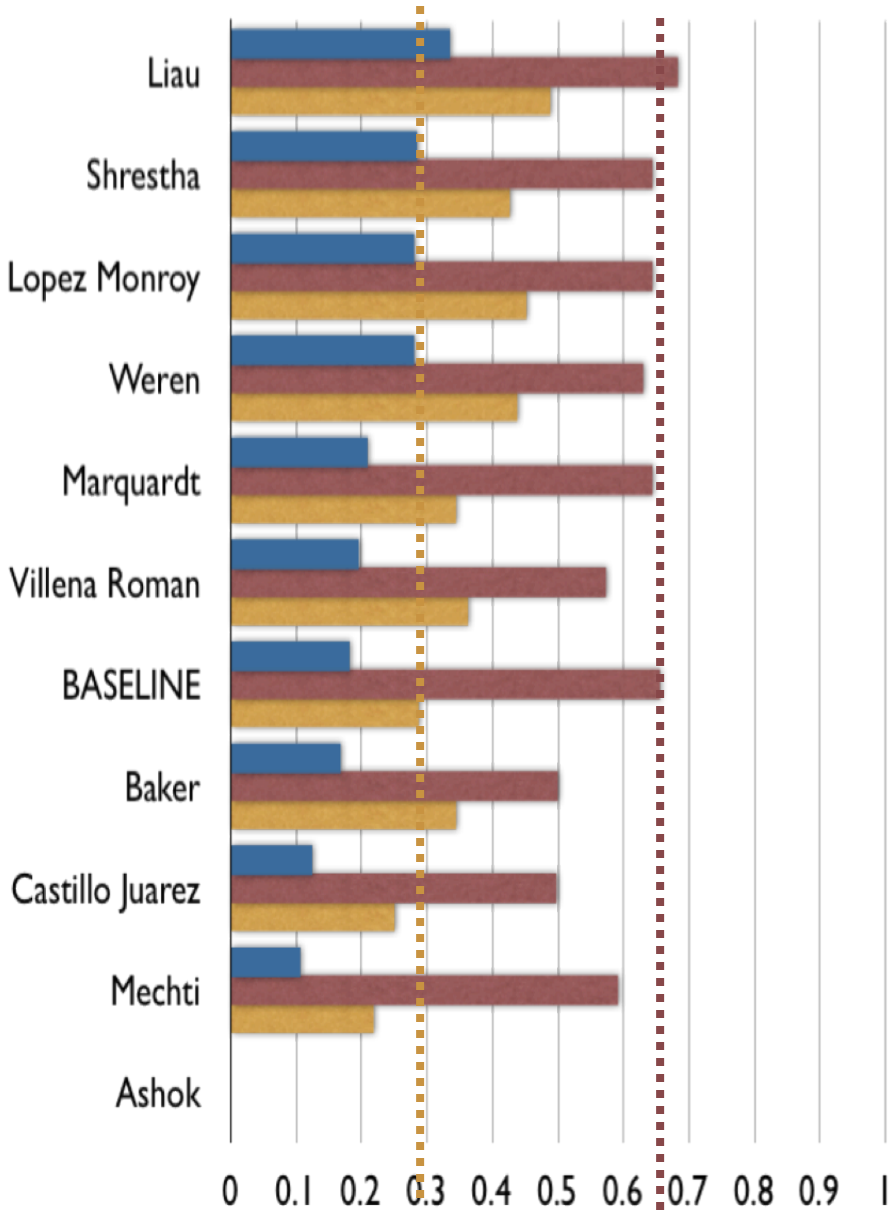


- **Social media + blogs + Twitter + reviews**
- **Age classes**: 18-24, 25-34, 35-49, 50-64, 65+

Results in social media

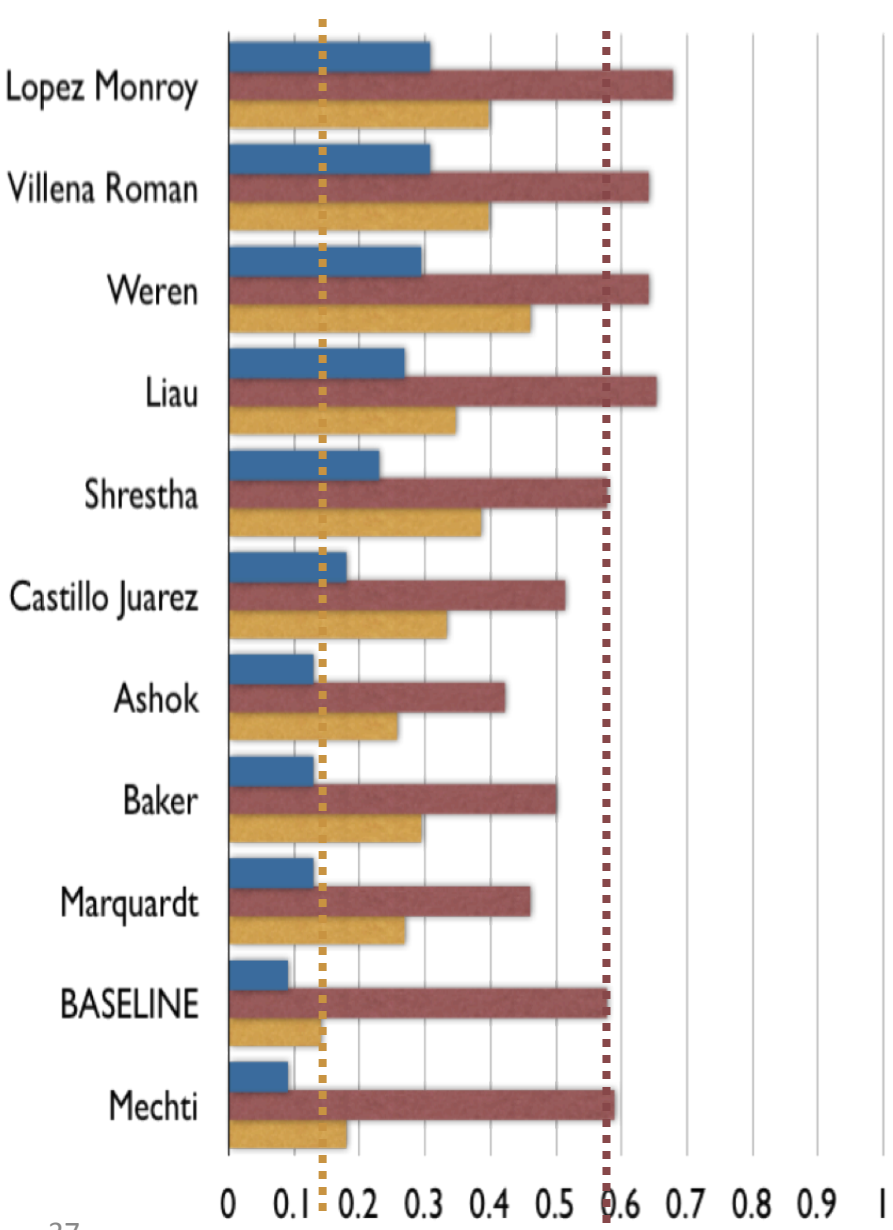


EN (joint, gender, age)

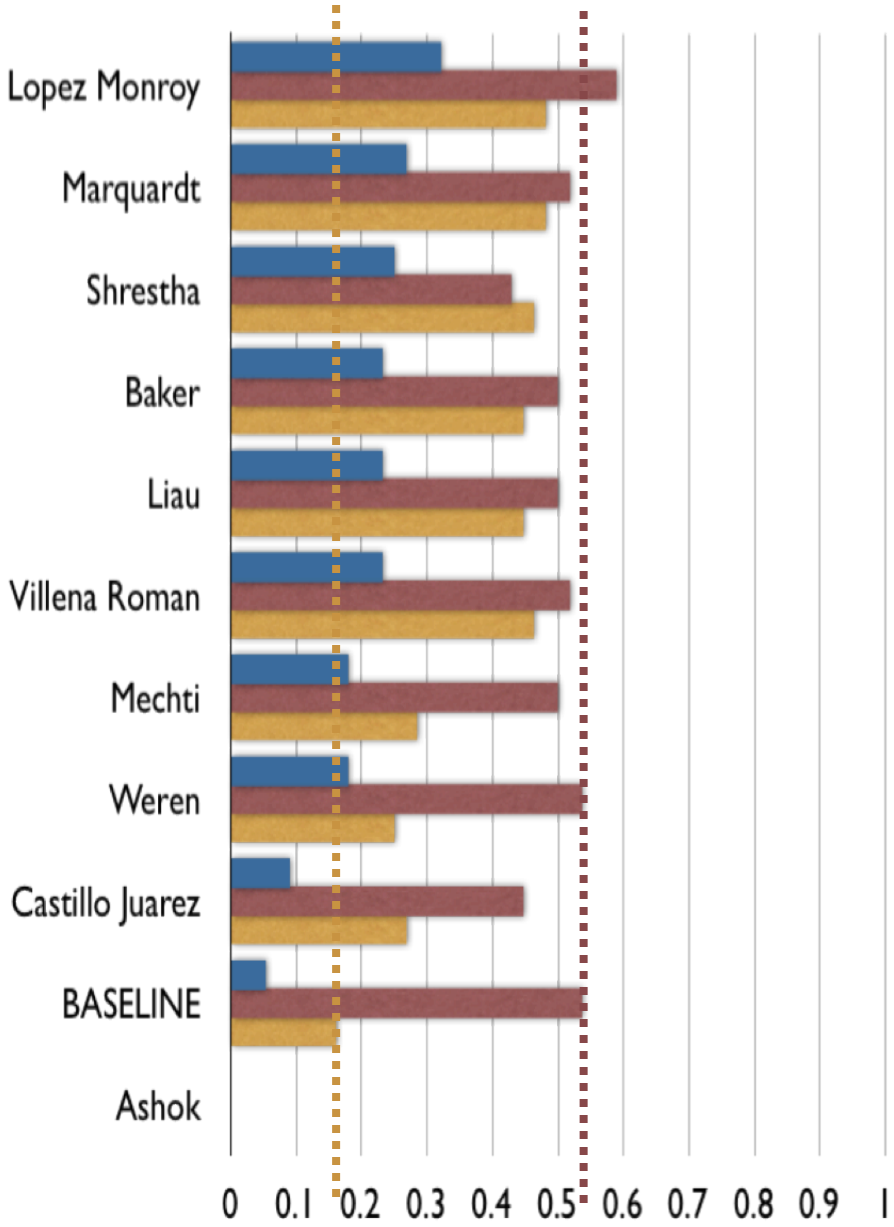


ES (joint, gender, age)

Results in blogs

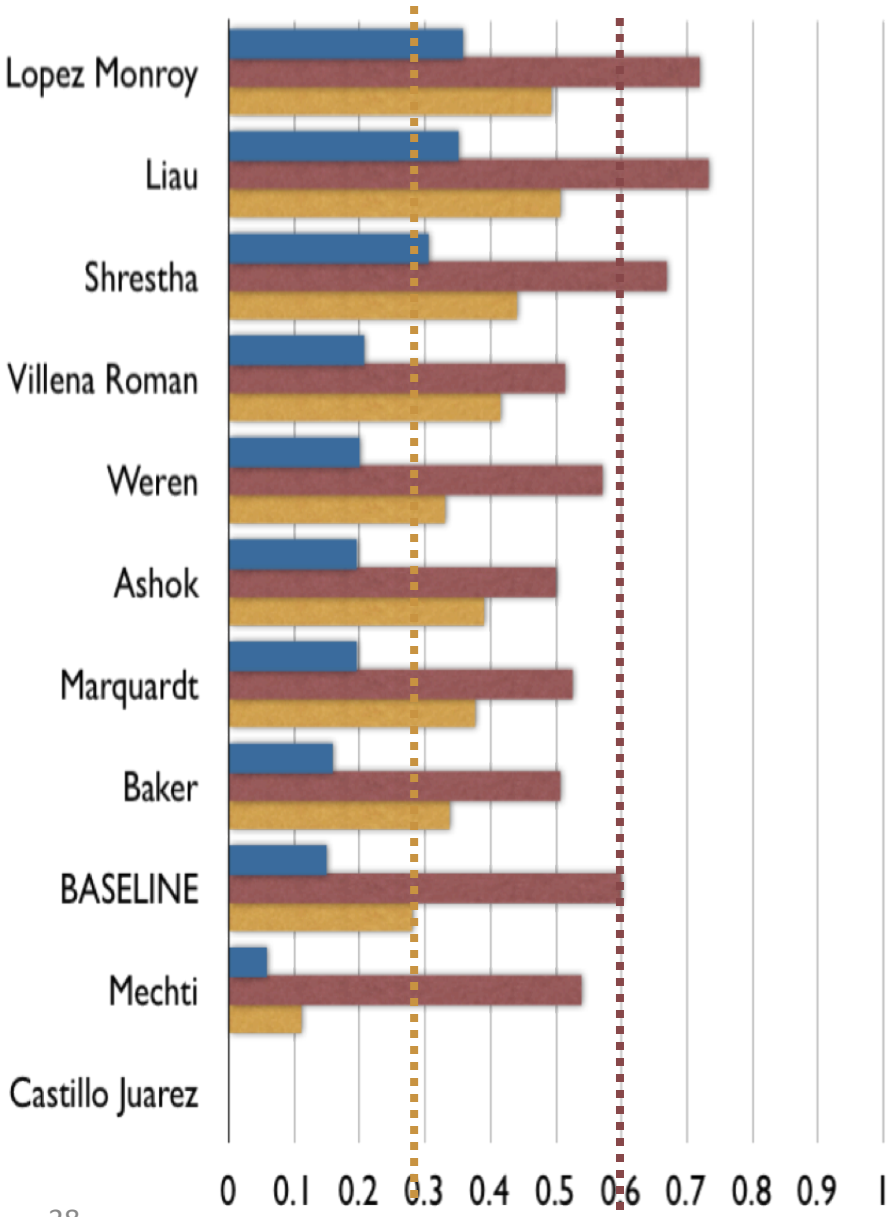


EN (joint, gender, age)

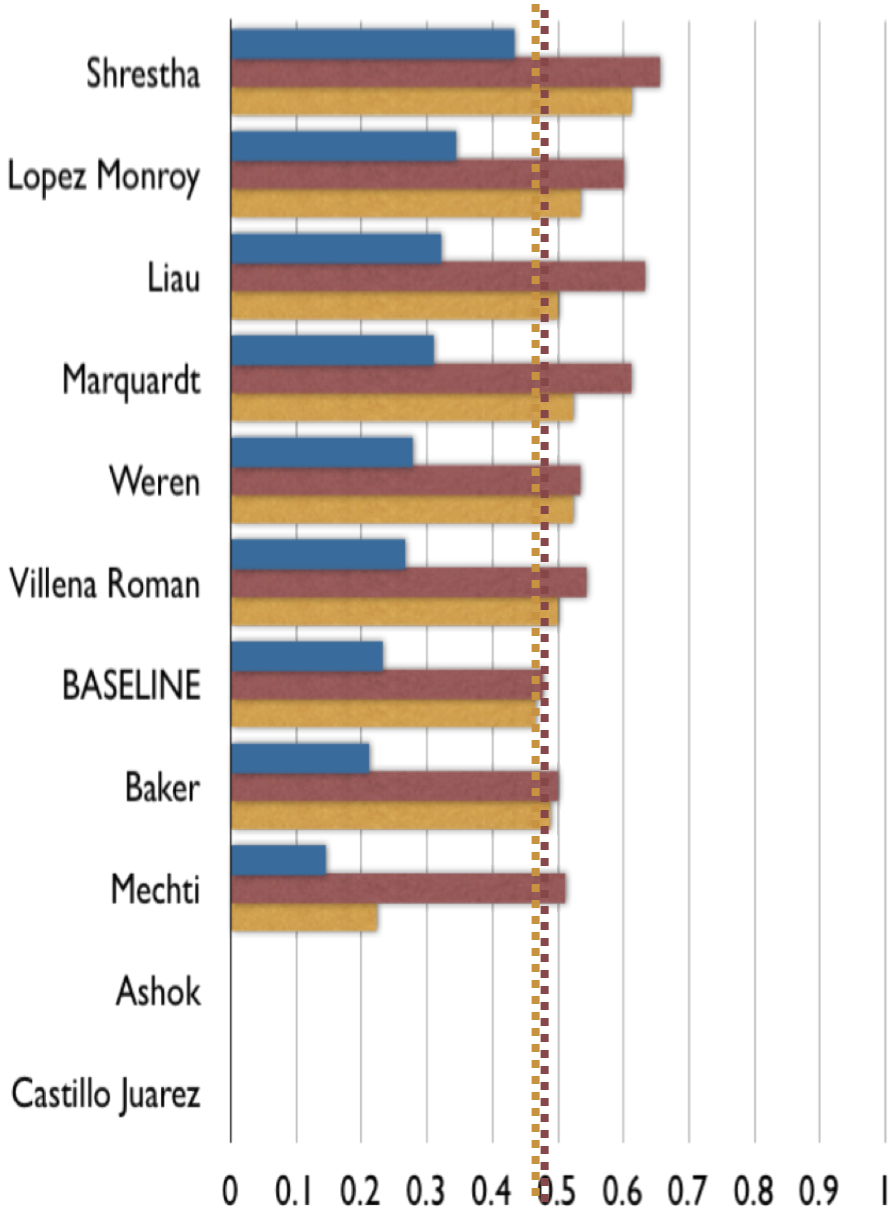


ES (joint, gender, age)

Results in Twitter

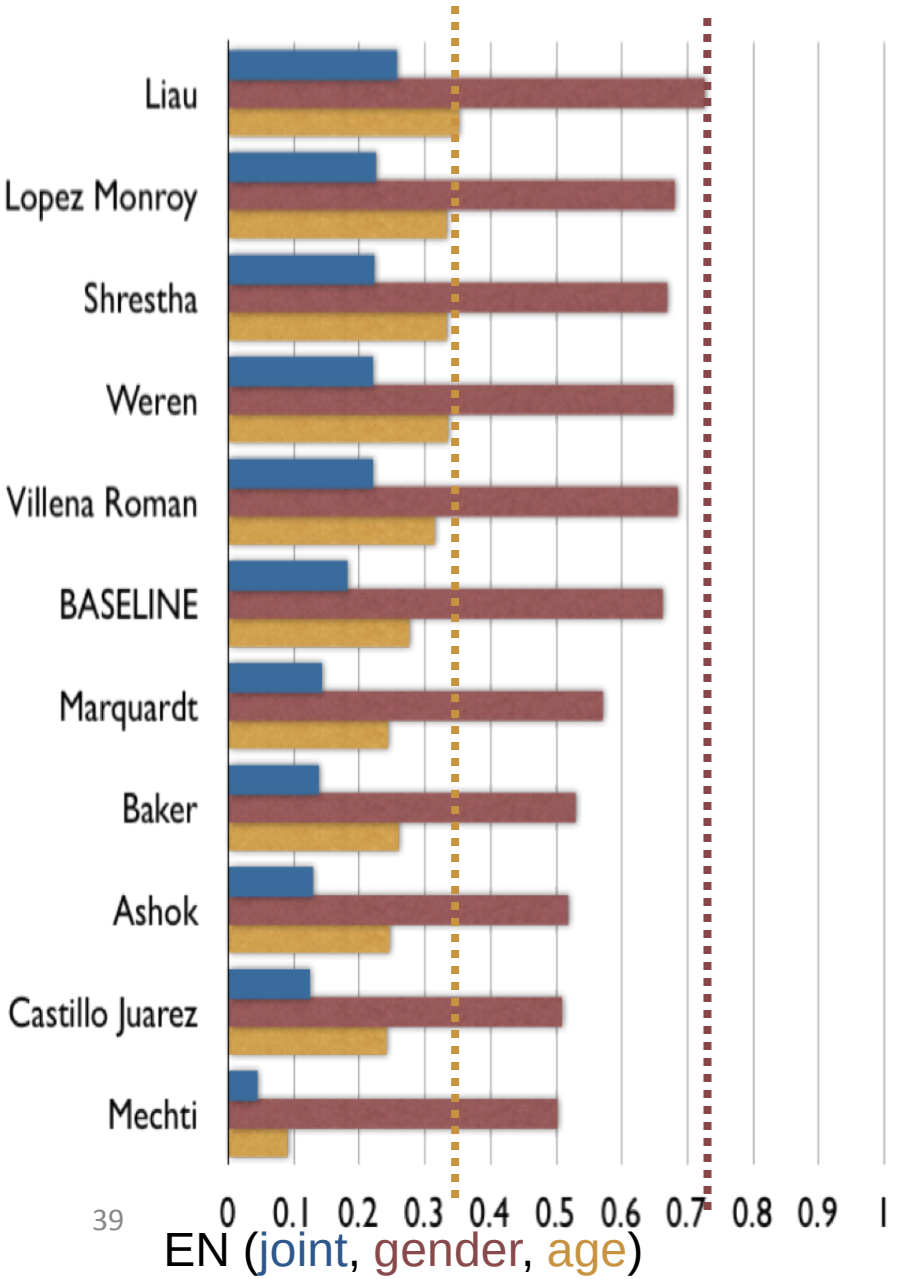


EN (joint, gender, age)



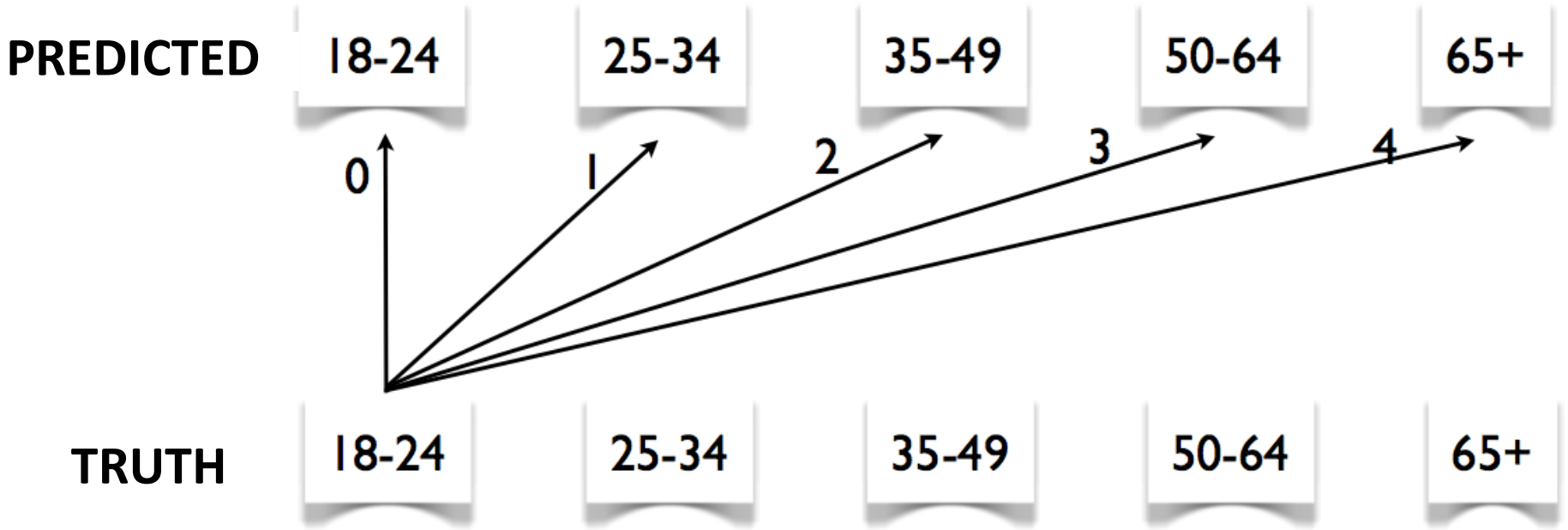
ES (joint, gender, age)

Results in reviews

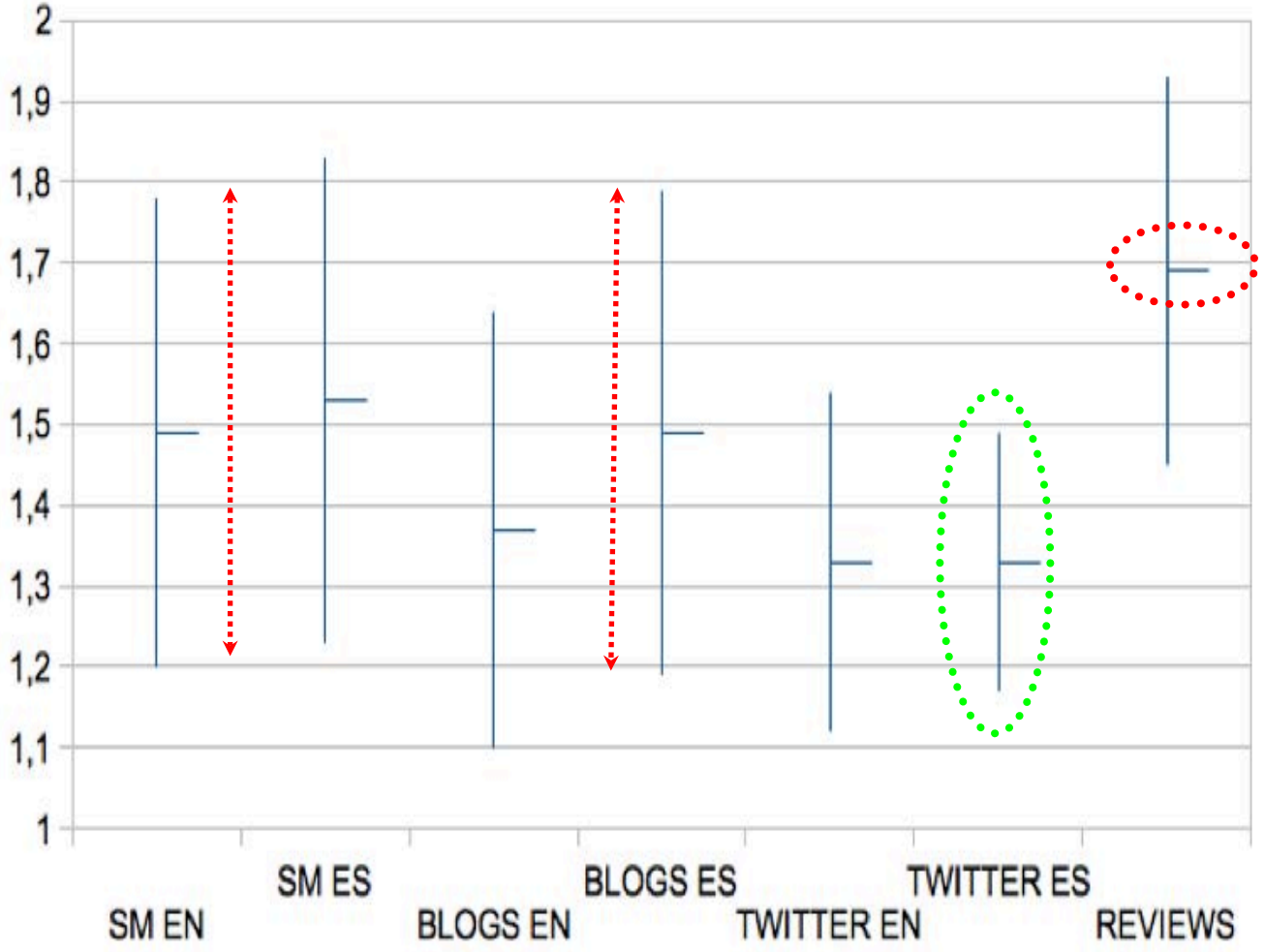


The problem of deceptive opinions

Distances in misclassified age



Distances in misclassified age



Twitter: more spontaneous way to communicate

Approaches: features

- Similar features than in 2013:

  content (**bag of words**, word n-grams) and stylistic

- frequency of words related to different psycholinguistic concepts, extracted from: LIWC and MRC psycholinguistic database

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

Author profiling: PAN@CLEF 2015

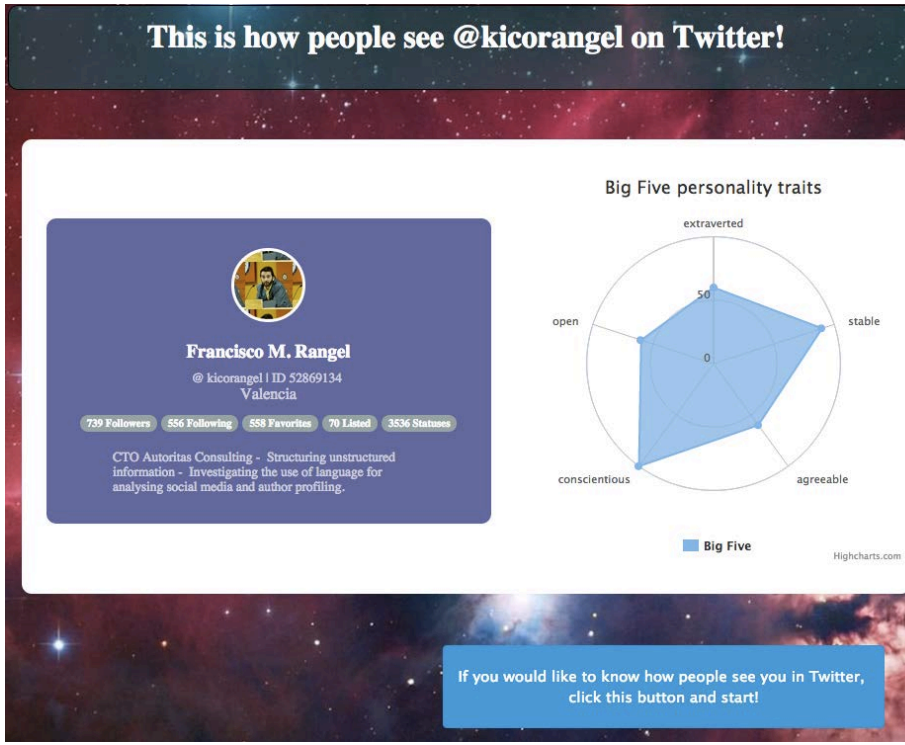
Gender, age, personality in Twitter

- Age classes:

  18-24, 25-34, 35-49, 50+
- Languages:

  English, Spanish, Italian, Dutch
- Teams submitting results: 22
- Best results (personality):

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



Rangel F., Celli F., Rosso P., Potthast M., Stein B., Daelemans W. Overview of the 3rd Author Profiling Task at PAN 2015. Notebook for PAN at CLEF 2016. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1391. 2015

Personality questionnaire

1. I am a reserved person
2. I trust other people
3. I tend to be lazy
4. I am generally relaxed, not stressed
5. I have few artistic interests
6. I am sociable
7. I tend to find fault with others
8. I do my job well
9. I get nervous easily
10. I have an active imagination

(answers from 1 to 5: http://mypersonality.autoritas.net/)

Previous shared tasks on personality recognition

http://mypersonality.org/wiki/doku.php?id=**wcpr**13

WCPR @ ACM Multimedia 2014

https://sites.google.com/site/wcprst/home/**wcpr**14

Big Five personality traits

Big Five personality traits (given a text, determine if the author is):

Open to new experiences
Conscientious: tends to be careful and scrupuolus
Extroverted: gets energy from being around people
Agreeable: prefers to agree with others
Neurotic: tends to worry about things

Accuracy results

Open/Closed (Friendly/Uncooperative)	66%
Conscientious (Organised/Careless)	65%
Extrovert/Shy	62%
Agreeable	60%
Neurotic/Stable	63%

Data: J. W. Pennebaker (students wrote essays and same students took personality assessment tests)

Some key features

- Openness
  - *consciousness, strange, thoughts, maybe, you*
  - *hope, feel, home, friends, football, team*
- Conscientiousness
  - *school, always, high, grades*
  - *damn, bad, hate, you, more*

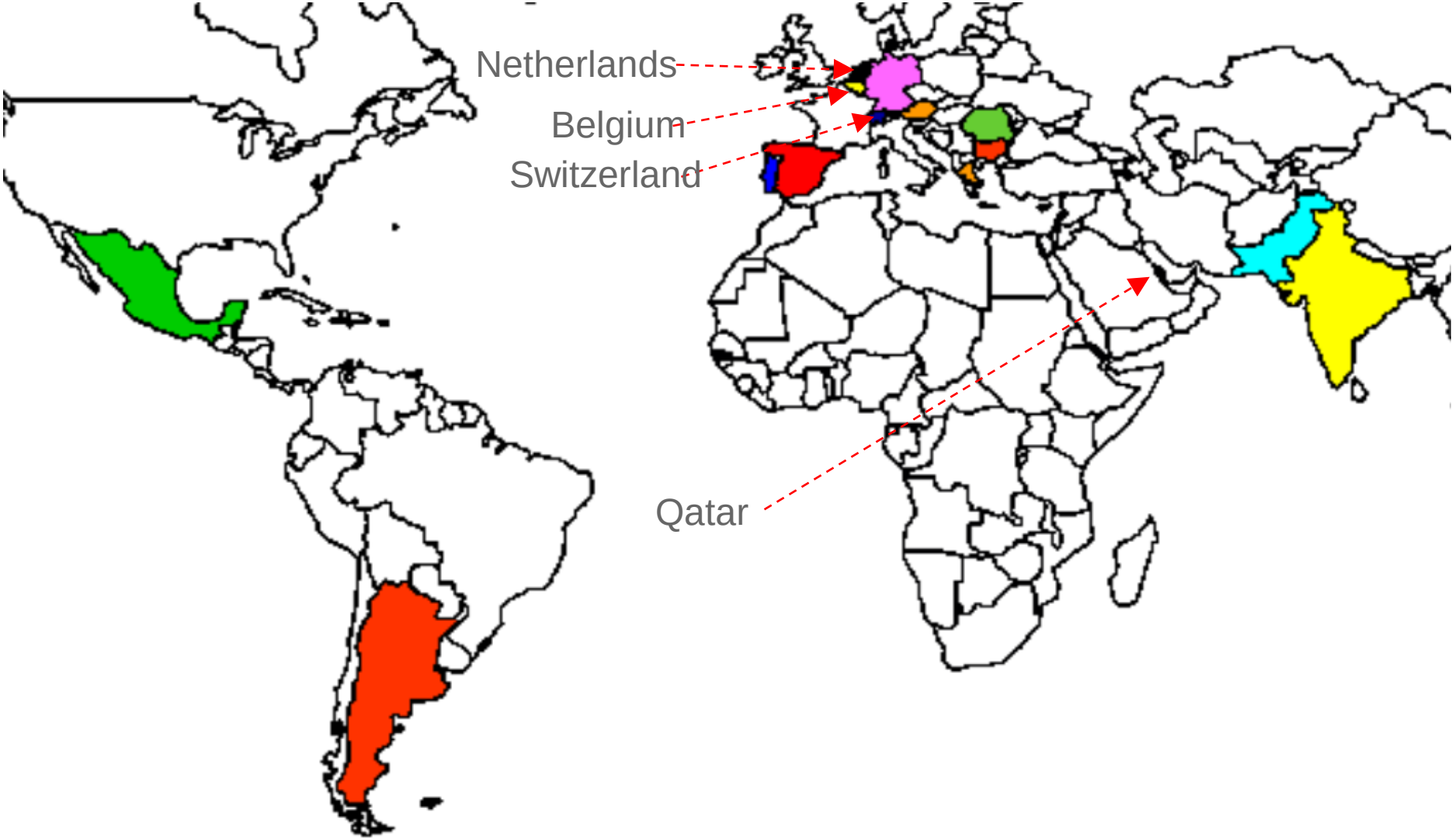
S. Argamon, S. Dawhle, M. Koppel and J. Pennebaker . Lexical Predictors of Personality Type, Proceedings of Classification Society of North America, St. Louis MI, 2005.

Author profiling: PAN@CLEF 2016

- **Cross-genre** gender and age : train **Twitter**
  test social media and blogs
- Age classes: 18-24, 25-34, 35-49, 35-49, 50-64, 65+
- Languages: English, Spanish, Dutch
- Teams submitting results: 22
- Training on Twitter data allowed to obtain competitive cross-genre results

Rangel F., Rosso M., Verhoeven B., Daelemans W., Potthast M., Stein B. Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. Notebook for PAN at CLEF 2016. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1609, pp. 750-784, 2016

Author profiling PAN @CLEF 2016



PR-SOCO : PAN @FIRE 2016

Personality Recognition in SOurce Code

- Big five personality traits from Java source codes
- 11 teams submitted 49 runs
- http://www.autoritas.es/prsoco/
- Best results for the openness trait

  in line with the results obtained on Twitter data

  at PAN @ CLEF 2015

Rangel F., González F., Restrepo-Calle F., Montes M., Rosso P. PAN at FIRE: Overview of the PR-SOCO Track on Personality Recognition in SOurce COde. Notebook for PAN at FIRE 2016. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1737, pp. 1-5. 2016

Author profiling: PAN @CLEF 2017

Gender and language variety in Twitter

LANGUAGE VARIETY			
ENGLISH	SPANISH	PORTUGUESE	ARABIC
• Australia • Canada • Great Britain • Ireland • New Zealand • United States	• Argentina • Chile • Colombia • Mexico • Peru • Spain • Venezuela	• Brazil • Portugal	• Egypt • Gulf • Levantine • Maghrebi

Rangel F., Rosso P., Potthast M., Stein B. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. Notebook for PAN at CLEF 2019. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1866. 2017

Corpus collection

- **Step 1:** Languages and varieties selection
- **Step 2:** Tweets per region retrieval

Language	Variety	City
Arabic	Egypgt	Cairo
	Gulf	Abu Dhabi, Doha, Kuwait, Manama, Mascate, Riyadh, Sana'a
	Levantine	Amman, Beirut, Damascus, Jerusalem
	Maghrebi	Algiers, Rabat, Tripoli, Tunis
English	Australia	Canberra, Sydney
	Canada	Toronto, Vancouver
	Great Britain	London, Edinburgh, Cardiff
	Ireland	Dublin
	New Zealand	Wellington
	United States	Washington
Portuguese	Brazil	Brasilia
	Portugal	Lisbon
Spanish	Argentina	Buenos Aires
	Chile	Santiago
	Colombia	Bogota
	Mexico	Mexico
	Peru	Lima
	Spain	Madrid
	Venezuela	Caracas

Corpus collection

- **Step 3:** Unique authors identification
- **Step 4:** Authors selection:
  - Tweets are not retweets
  - Tweets are written in the corresponding language
- **Step 5:** Language variety annotation:
  - 80% of tweet meta-data coincide with:
    - Geotagging
    - Toponyms of the region
- **Step 6:** Gender annotation:
  - Automatically: dictionary of proper nouns
  - Manually: review

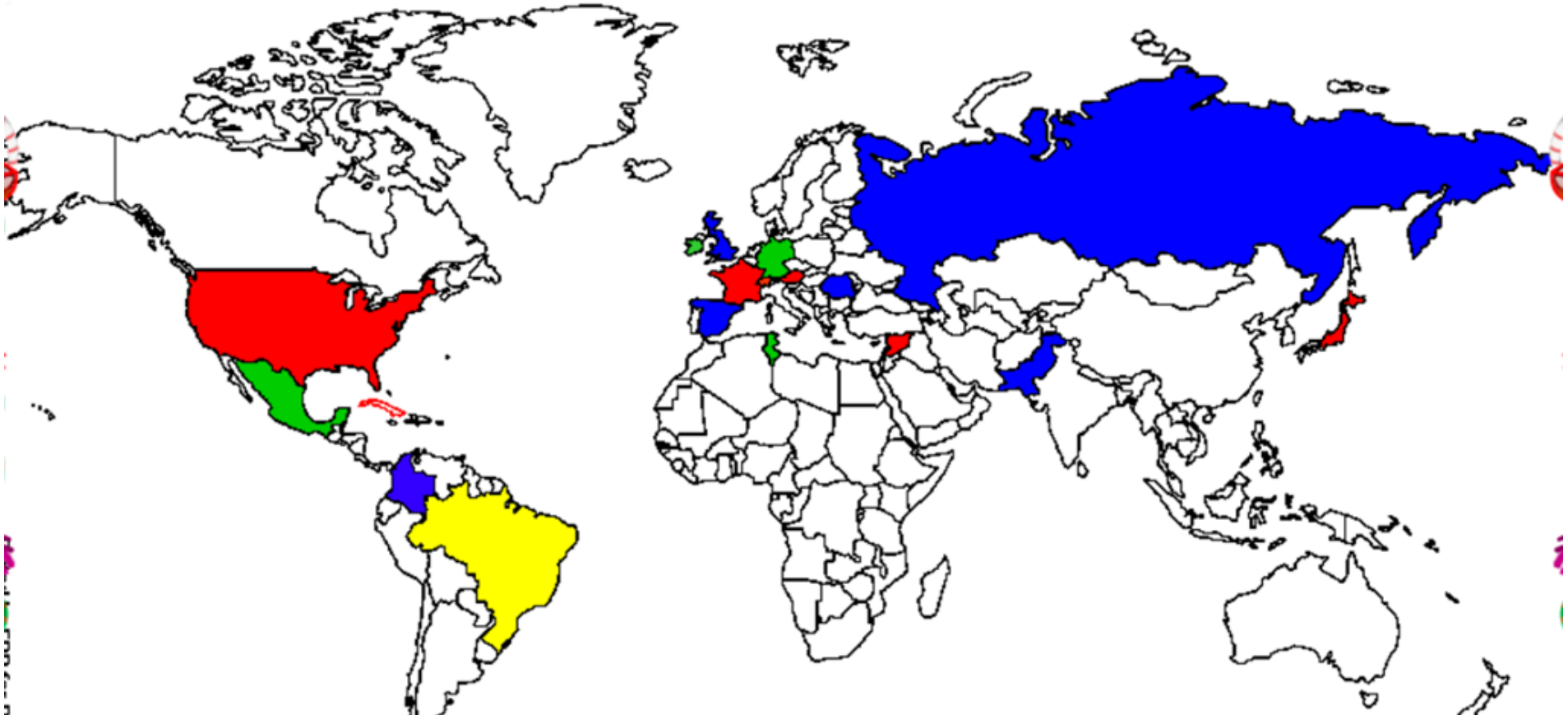
Corpus collection

- **Step 7:** Corpus construction:
  - 500 authors per variety and gender
    - **300 for training, 200 for test** (to avoid overfitting)
  - 100 tweets per author

(AR) Arabic	(EN) English	(ES) Spanish	(PT) Portuguese
Egypt	Australia	Argentina	Brazil
Gulf	Canada	Chile	Portugal
Levantine	Great Britain	Colombia	
Maghrebi	Ireland	Mexico	
	New Zealand	Peru	
	United States	Spain	
		Venezuela	
4,000	6,000	7,000	2,000

Participants

22 teams submitting results: Brazil, Colombia, Cuba, France, Germany, Ireland, Japan, Mexico, Pakistan, Romania, Russia, Slovenia, Spain, Syria, Switzerland, The Netherlands, Tunisia, UK, USA... **ITALY: where are you??**



Preprocessing

HTML cleaning to obtain plain text	Khan. Martinc *et al.*; Ribeiro-Oliveira & Ferreira
Punctuation signs	Ribeiro-Oliveira & Ferreira; Martinc *et al.;* Schaetti
Stop words	Kheng *et al.*; Martinc *et al.*
Lowercase	Franco-Salvador *et al.*; Kheng *et al.*; Kodiyan *et al.*; Miura *et al.*
Remove short tweets	Kheng *et al.*
Twitter specific components: hashtags, urls, mentions and RTs	Franco-Salvador *et al.*; Adame *et al.*; Kheng *et al.*; Kodiyan *et al.*; Markov *et al.*; Miura *et al.*; Ribeiro-Oliveira & Ferreira; Schaetti
Out-of-vocabulary words	Schaetti
Expand contractions	Adame *et al.*

Features

Stylistic features: - Ratios of links - Hashtag or user mentions - Character flooding - Emoticons / laugher expressions - Domain names	Alrifai *et al.*; Ribeiro-Oliveira & Ferreira; Martinc *et al.*; Adame *et al.*; Markov *et al.*
Emotional features: ● Emotions ● Appraisal ● Admiration ● Pos/neg emoticons ● Sentiment words ● ...	Adame *et al.*; Martinc *et al.*
Specific lists of words, most discriminant words, ..	Martinc *et al.*; Kocher & Savoy; Khan

Features

N-gram models	Martinc *et al.;*, Alrifai *et al.*; Kheng *et al.*; Markov *et al.*; Ribeiro-Oliveira & Ferreira; Ogaltsov & Romanov; Schaetti; Ciobanu *et al.*
Bag-of-words	Adame *et al.*; Tellez *et al.*
Tf-idf n-grams	Poulston *et al.;* Schaetti; Basile *et al.*
Latent Semantic Analysis	Kheng *et al.*
Second order representation	Pastor e*t al.*
Word embeddings	Ignatov *et al.;* Kodiyan *et al.*; Sierra *et al.;* Poulston *et al.;* Miura *et al.*
Character embeddings	Franco-Salvador *et al.*; Miura *et al.*

Methods

Logistic regression	Ignatov *et al.;* Martinc *et al.;* Poulston *et al.;* Ogaltsov & Romanov
SVM	Alrifai *et al.*; Kheng *et al.*; Pastor *et al.;* Markov *et al.*; Tellez *et al.;* Basile *et al.;* Ribeiro-Oliveira & Ferreira; Ciobanu *et al.*
Naive Bayes	Kheng *et al.*
Recurrent Neural Networks	Kodiyan *et al.*; Miura *et al.*
Convolutional Neural Networks	Schaetti; Sierra *et al.;* Miura *et al.*
Deep Averaging Networks	Franco-Salvador *et al.*

Baselines

- **BASELINE-stat:** A statistical baseline that emulates random choice
- **BASELINE-bow:**
  - Documents represented as bag-of-words
  - **The 1,000 most common words in the training set**
  - **Weighted by absolute frequency**
  - **Preprocess: lowercase, removal of punctuation signs and numbers, removal of stopwords**
- **BASELINE-LDR:**
  - Documents represented by the probability distribution of occurrence of their words in the different classes
  - Each word is weighted depending on its probability of belonging to each class
  - The distribution of weights for a given document should be closer to the weights of its corresponding class

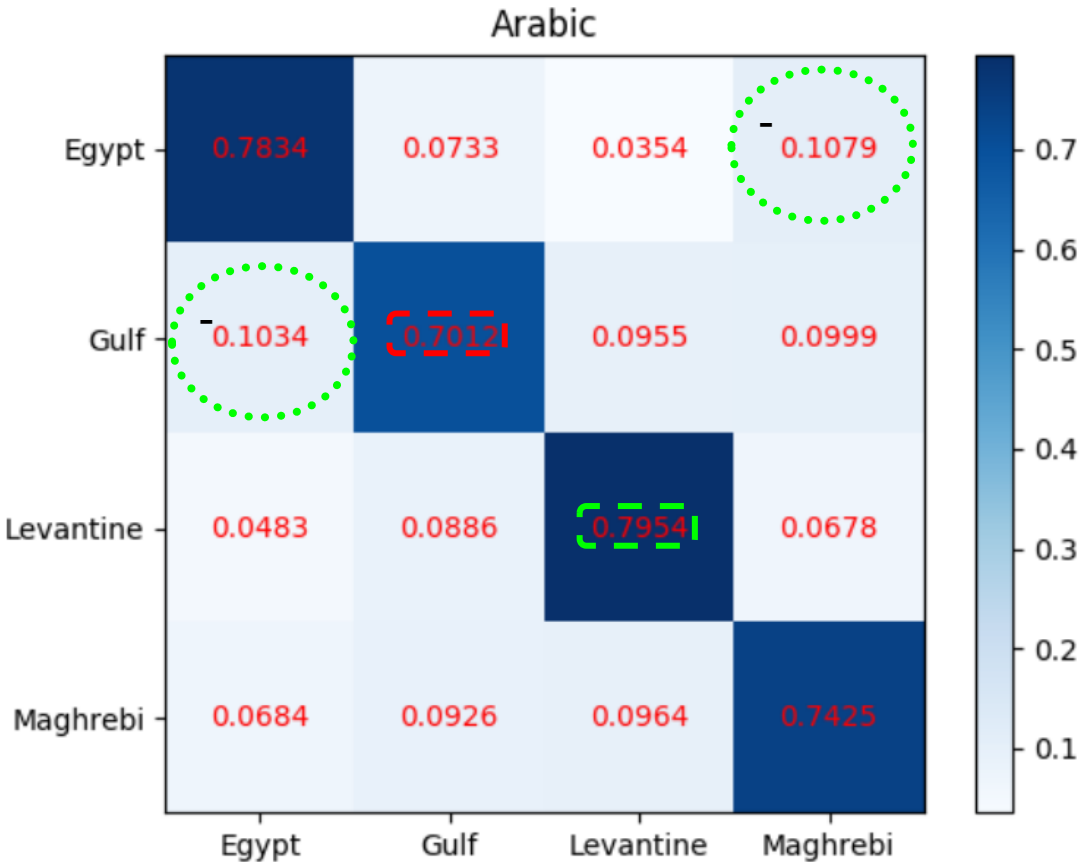
Results: gender

Ranking	Team	Arabic	English	Portuguese	Spanish	Average
1	Basile et al.	0.8006	0.8233	0.8450	0.8321	0.8253
2	Martinc et al.	0.8031	0.8071	0.8600	0.8193	0.8224
3	Miura et al.	0.7644	0.8046	0.8700	0.8118	0.8127
4	Tellez et al.	0.7838	0.8054	0.8538	0.7957	0.8097
5	Lopez-Monroy et al.	0.7763	0.8171	0.8238	0.8014	0.8047
6	Poulston et al.	0.7738	0.7829	0.8388	0.7939	0.7974
7	Markov et al.	0.7719	0.8133	0.7863	0.8114	0.7957
8	Ogaltsov & Romanov	0.7213	0.7875	0.7988	0.7600	0.7669
9	Franco-Salvador et al.	0.7300	0.7958	0.7688	0.7721	0.7667
10	Sierra et al.	0.6819	0.7821	0.8225	0.7700	0.7641
11	Kodiyan et al.	0.7150	0.7888	0.7813	0.7271	0.7531
12	Ciobanu et al.	0.7131	0.7642	0.7713	0.7529	0.7504
13	Ganesh	0.6794	0.7829	0.7538	0.7207	0.7342
	LDR-baseline	0.7044	0.7220	0.7863	0.7171	0.7325
14	Schaetti	0.6769	0.7483	0.7425	0.7150	0.7207
15	Kocher & Savoy	0.6913	0.7163	0.7788	0.6846	0.7178
16	Kheng et al.	0.6856	0.7546	0.6638	0.6968	0.7002
17	Ignatov et al.	0.6425	0.7446	0.6850	0.6946	0.6917
	BOW-baseline	0.5300	0.7075	0.7812	0.6864	0.6763
18	Khan	0.5863	0.6692	0.6100	0.6354	0.6252
	STAT-baseline	0.5000	0.5000	0.5000	0.5000	0.5000
19	Ribeiro-Oliveira et al.	0.7013	-	0.7650	-	0.3666
20	Alrifai et al.	0.7225	-	-	-	0.1806
21	Bouzazi	-	0.6121	-	-	0.1530
22	Adame et al.	-	0.5413	-	-	0.1353

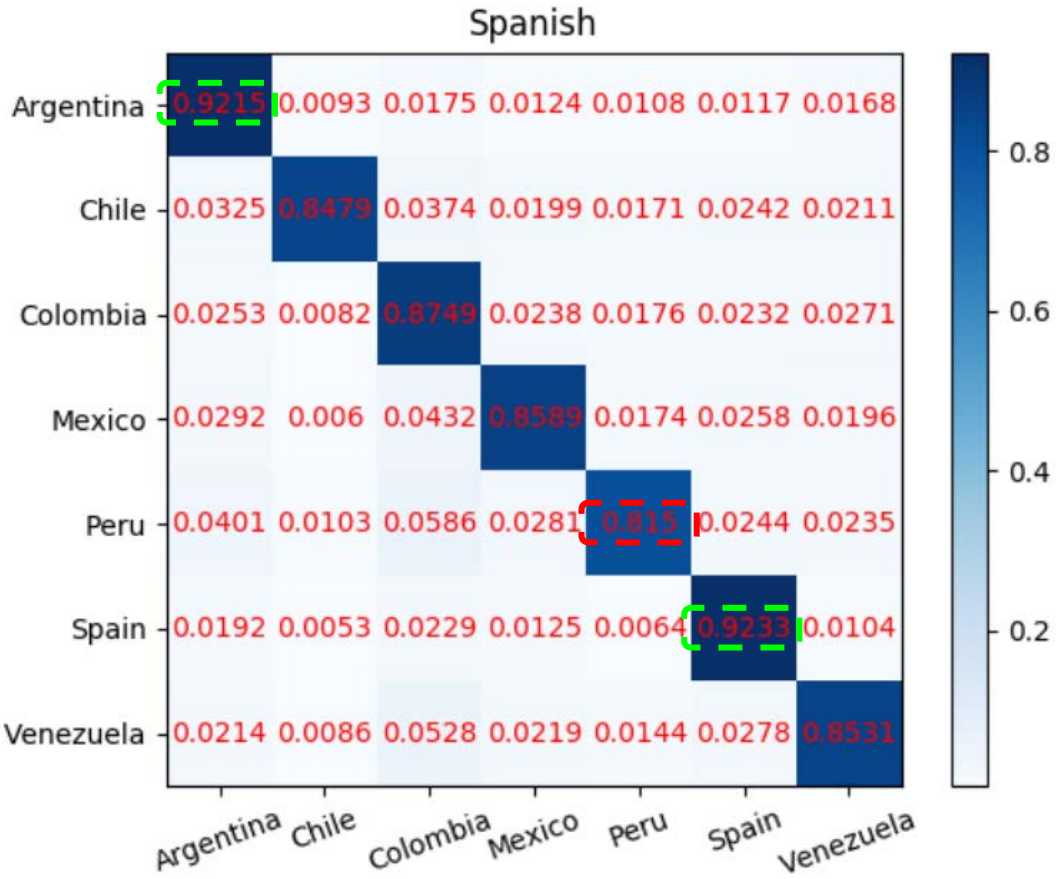
Results: language variety

Ranking	Team	Arabic	English	Portuguese	Spanish	Average
	LDR-baseline	0.8250	0.8996	0.9875	0.9625	0.9187
1	Basile et al.	0.8313	0.8988	0.9813	0.9621	0.9184
2	Tellez et al.	0.8275	0.9004	0.9850	0.9554	0.9171
3	Martinc et al.	0.8288	0.8688	0.9838	0.9525	0.9085
4	Markov et al.	0.8169	0.8767	0.9850	0.9439	0.9056
5	Lopez-Monroy et al.	0.8119	0.8567	0.9825	0.9432	0.8986
6	Miura et al.	0.8125	0.8717	0.9813	0.9271	0.8982
7	Sierra et al.	0.7950	0.8392	0.9850	0.9450	0.8911
8	Schaetti	0.8131	0.8150	0.9838	0.9336	0.8864
9	Poulston et al.	0.7975	0.8038	0.9763	0.9368	0.8786
10	Ogaltsov & Romanov	0.7556	0.8092	0.9725	0.8989	0.8591
11	Ciobanu et al.	0.7569	0.7746	0.9788	0.8993	0.8524
12	Kodiyan et al.	0.7688	0.7908	0.9350	0.9143	0.8522
13	Kheng et al.	0.7544	0.7588	0.9750	0.9168	0.8513
14	Franco-Salvador et al.	0.7656	0.7588	0.9788	0.9000	0.8508
15	Kocher & Savoy	0.7188	0.6521	0.9725	0.7211	0.7661
16	Ganesh	0.7144	0.6021	0.9650	0.7689	0.7626
17	Ignatov et al.	0.4488	0.5813	0.9763	0.8032	0.7024
	BOW-baseline	0.3394	0.6592	0.9712	0.7929	0.6907
18	Khan	0.5844	0.2779	0.9063	0.3496	0.5296
19	Ribeiro-Oliveira et al.	0.6713	-	0.9850	-	0.4141
	STAT-baseline	0.2500	0.1667	0.5000	0.1429	0.2649
20	Alrifai et al.	0.7550	-	-	-	0.1888
21	Bouzazi	-	0.3725	-	-	0.0931
22	Adame et al.	-	0.1904	-	-	0.0476

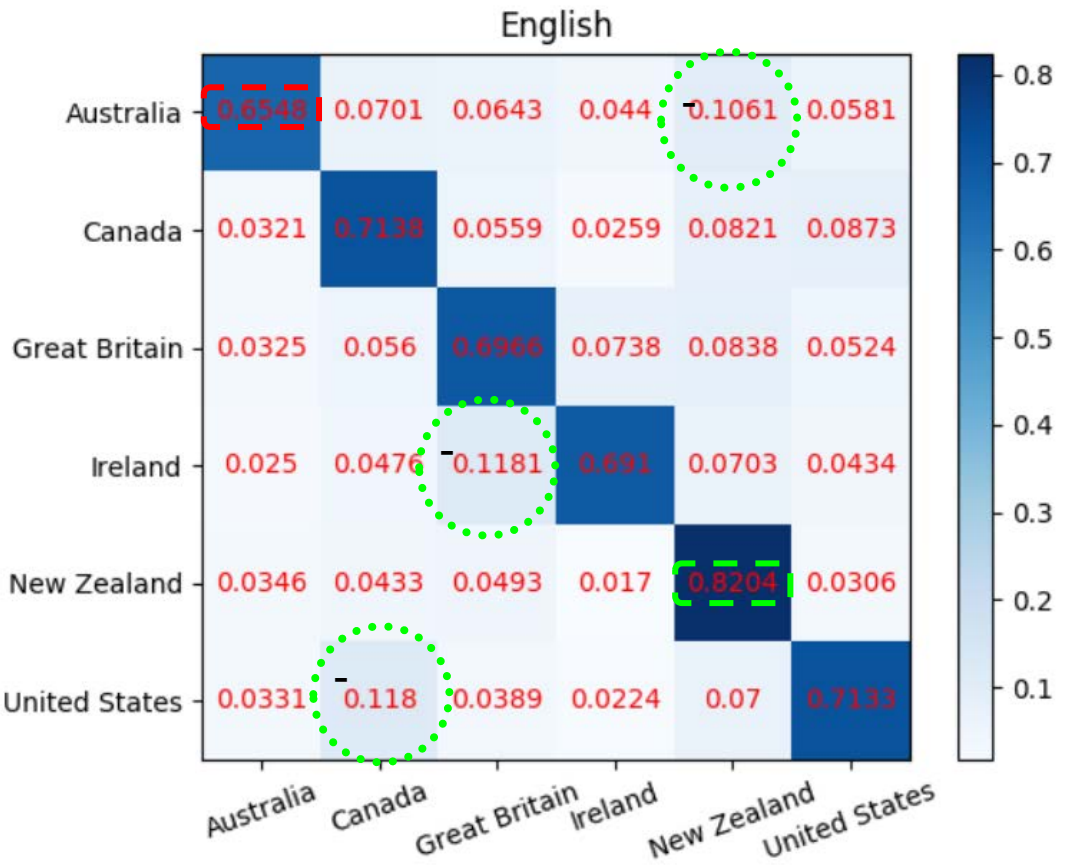
Confusion among AR varieties



Confusion among ES varieties



Confusion among EN varieties



EN coarse vs. fine grain

- **American:** United States + Canada
- **European:** Great Britain + Ireland
- **Oceanic:** New Zealand + Australia

Ranking	Team	Coarse-Grained	Fine-Grained	Difference
1	Basile et al.	0.9429	0.8988	0.0441
2	Tellez et al.	0.9379	**0.9004**	**0.0375**
3	Markov et al.	0.9292	0.8767	0.0525
4	Miura et al.	0.9279	0.8717	0.0562
5	Martinc et al.	0.9238	0.8688	0.0550
6	Lopez-Monroy et al.	0.9167	0.8567	0.0600
7	Sierra et al.	0.9004	0.8392	0.0612
8	Schaetti	0.8863	0.8150	0.0713
9	Ogaltsov & Romanov	0.8754	0.8092	0.0662
10	Poulston et al.	0.8746	0.8038	0.0708
11	Kodiyan et al.	0.8663	0.7908	0.0755
12	Franco-Salvador et al.	0.8654	0.7588	0.1066
13	Ciobanu et al.	0.8504	0.7746	0.0758
14	Kheng et al.	0.8388	0.7588	0.0800
15	Kocher & Savoy	0.7696	0.6521	0.1175
16	Ignatov et al.	0.7296	0.5813	0.1483
17	Ganesh	0.7238	0.6021	0.1217
18	Bouzazi	0.5217	0.3725	0.1492
19	Khan	0.4533	0.2779	**0.1754**
20	Adame et al.	0.3583	0.1904	0.1679

Author profiling: industry @ PAN

Organisation

autoritas
nuevas ideas, nuevas soluciones

Participants

DATYS
SOLUCIONES TECNOLÓGICAS

FUJI Xerox

ANCHORMEN

KNOW Center

OTS
OPTICAL TECH & SUPPORT

iSG

Bitdefender

Sponsors

meaning cloud

symanto GROUP

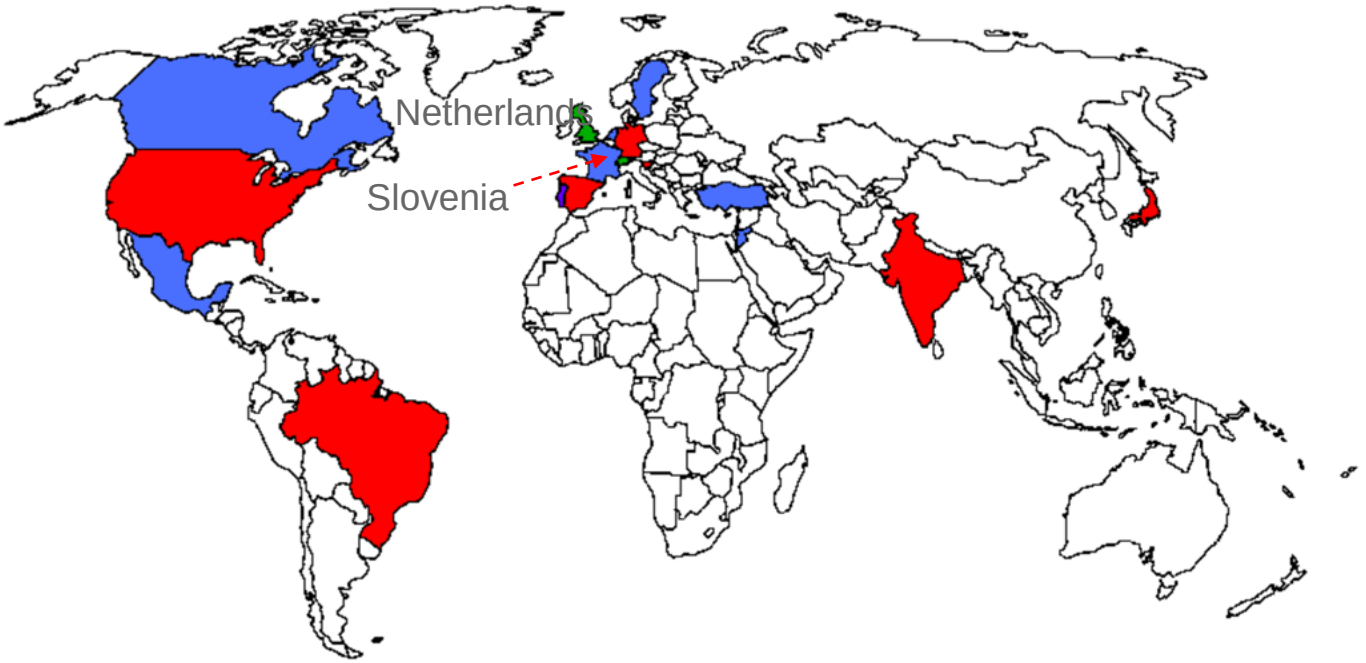
Adobe

Author profiling: participation @ PAN

	PARTICIPANTS	COUNTRIES
PAN-AP 2013	21	16
PAN-AP 2014	10	8
PAN-AP 2015	22	13
PAN-AP 2016	22	15
PAN-AP 2017	22	19
PAN-AP 2018	23	17

Author profiling: PAN@CLEF 2018

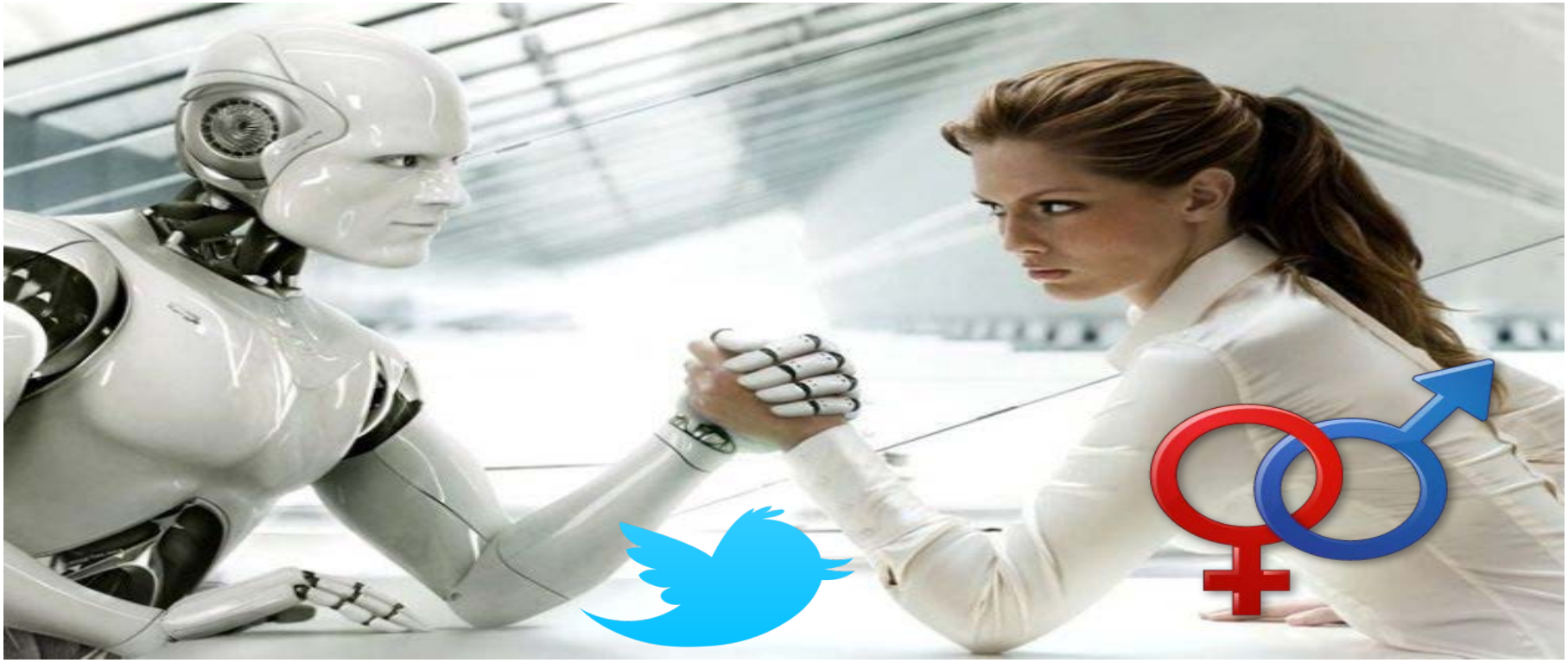
- **Multimodal** gender identification in Twitter
- Languages: Arabic, English, Spanish



Rangel F., Rosso M., Montes y Gómez M., Potthast M., Stein B. Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. Notebook for PAN at CLEF 2018. CEUR Workshop Proceedings. CEUR-WS.org, vol. 2125, 2018

Author profiling: PAN@CLEF 2019

Bots and gender profiling



2019: the year of Italy in author profiling??

EmoGraph based discourse analysis

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

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

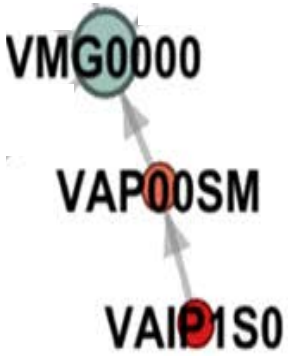
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.

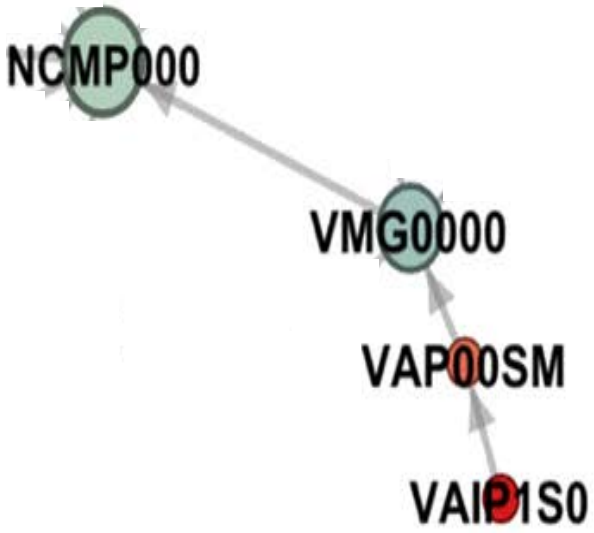
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



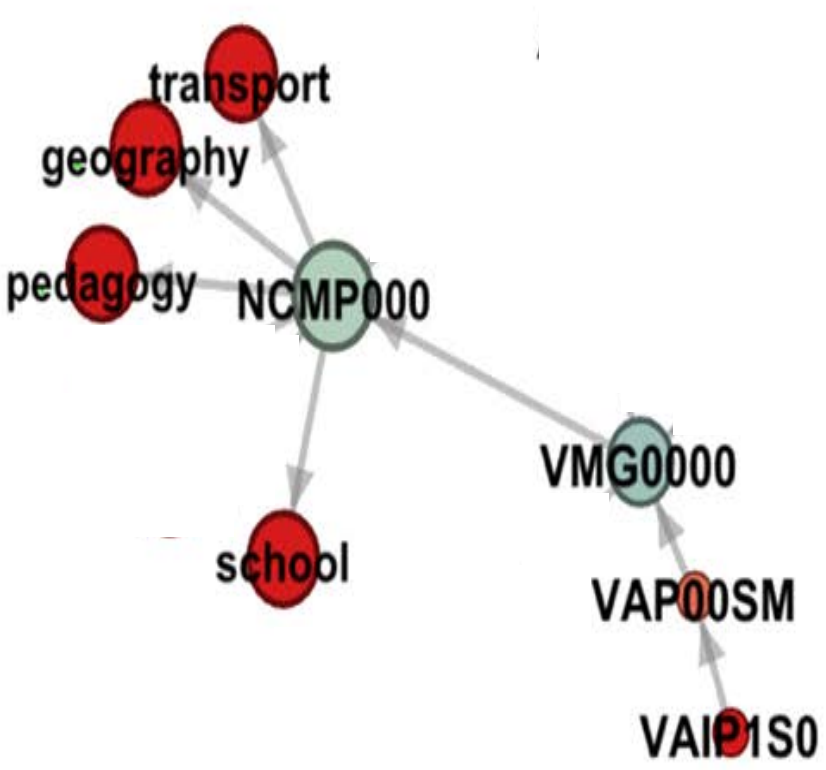
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online **courses** about valuable subjects that (I) enjoy studying and that might help me to speak in public.



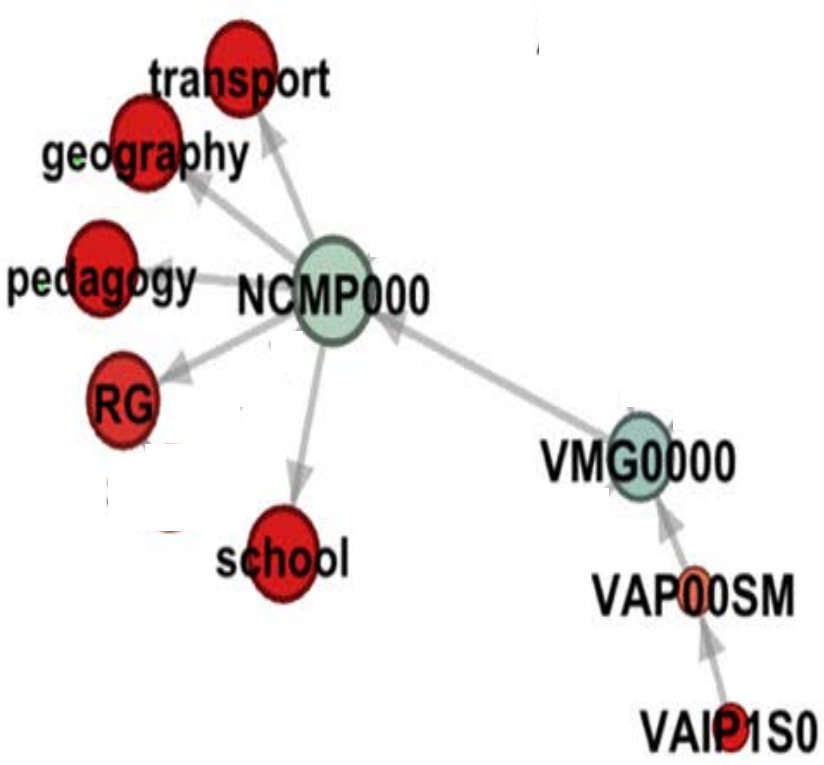
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online **courses** about valuable subjects that (I) enjoy studying and that might help me to speak in public.



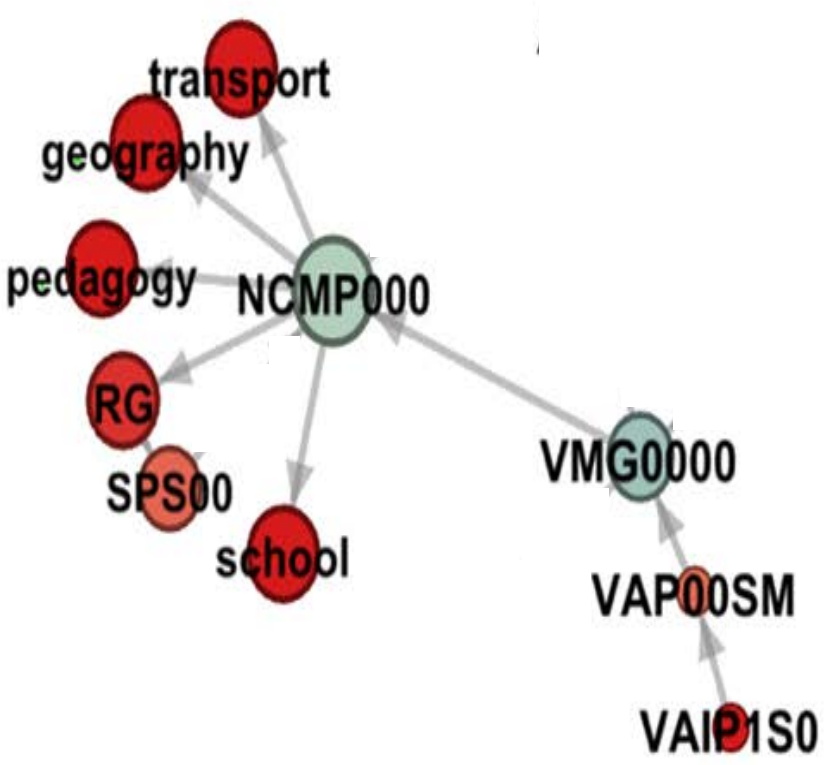
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



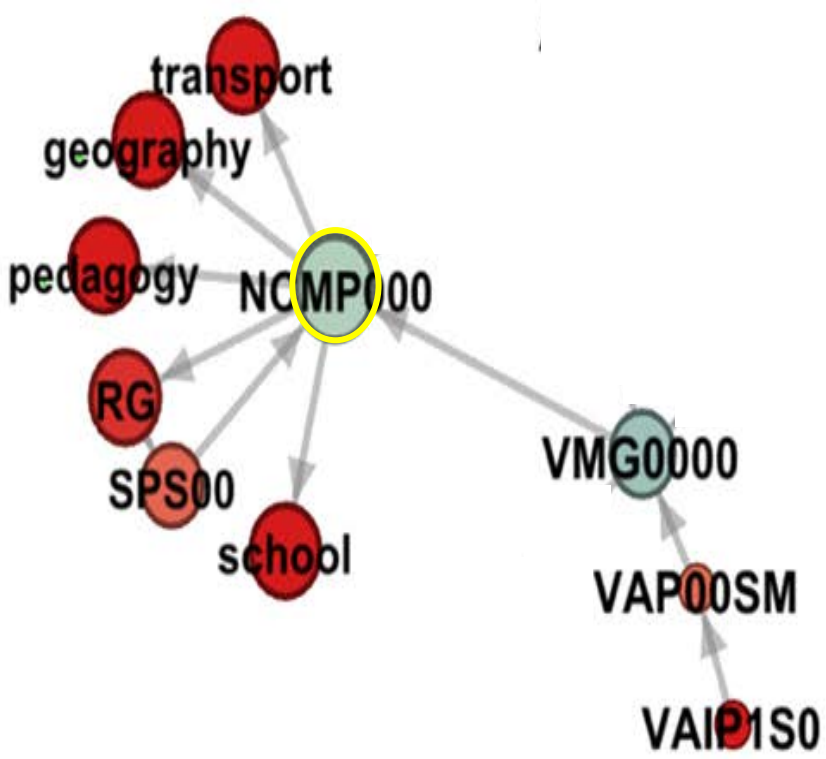
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



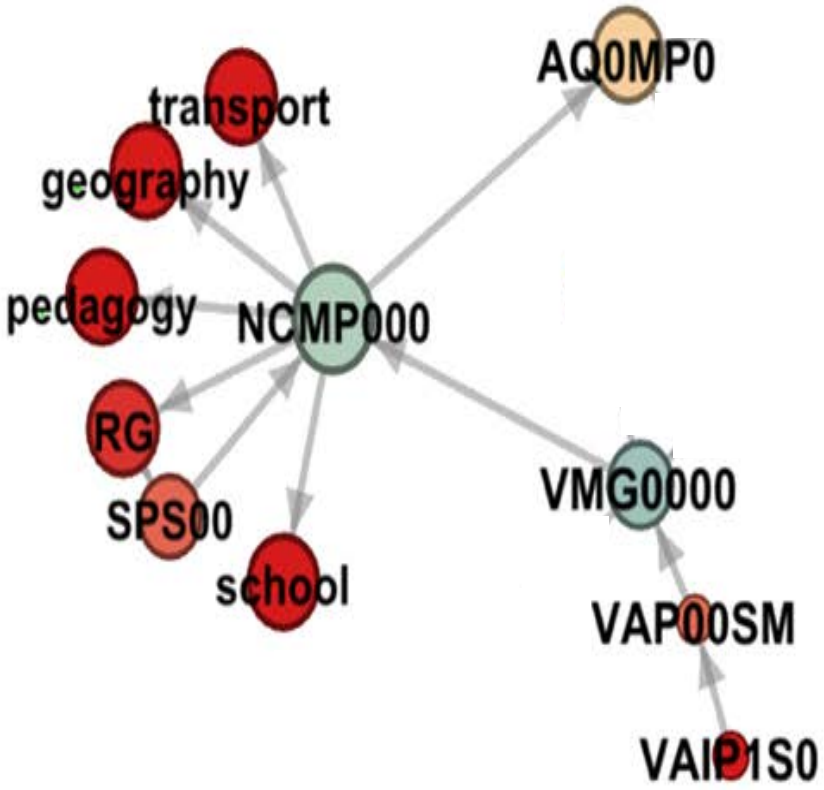
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable **subjects** that (I) enjoy studying and that might help me to speak in public.



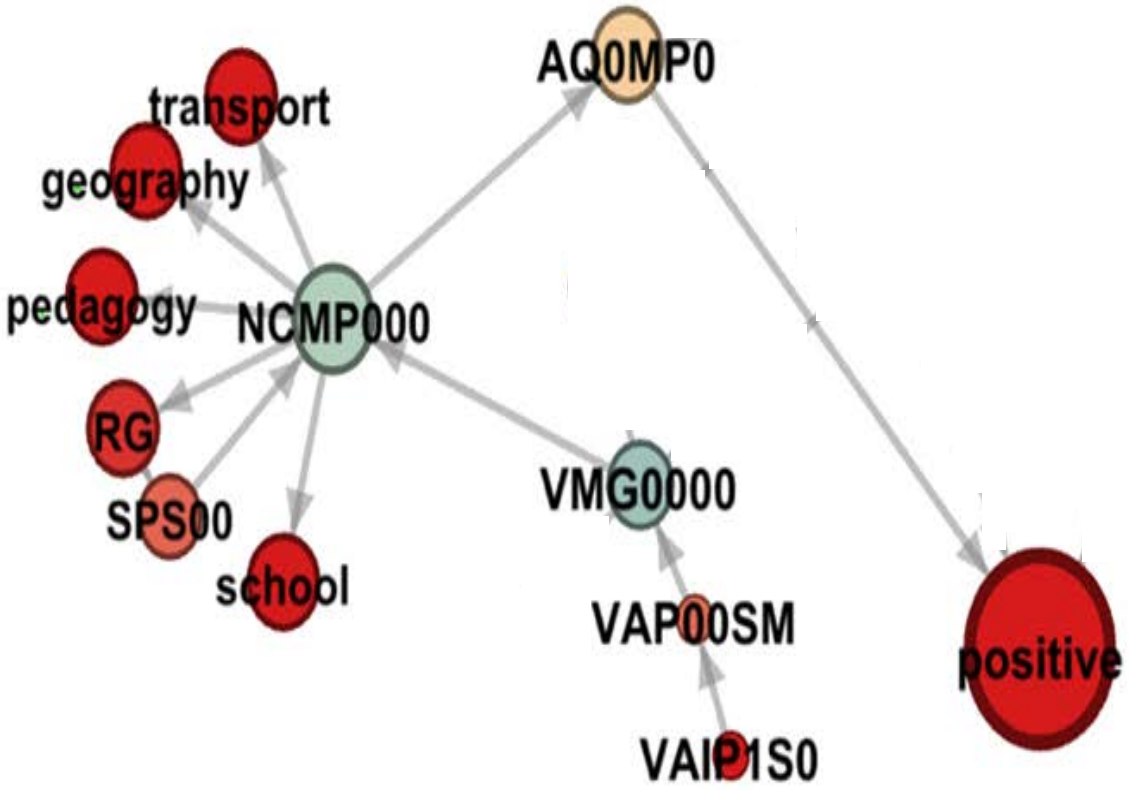
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



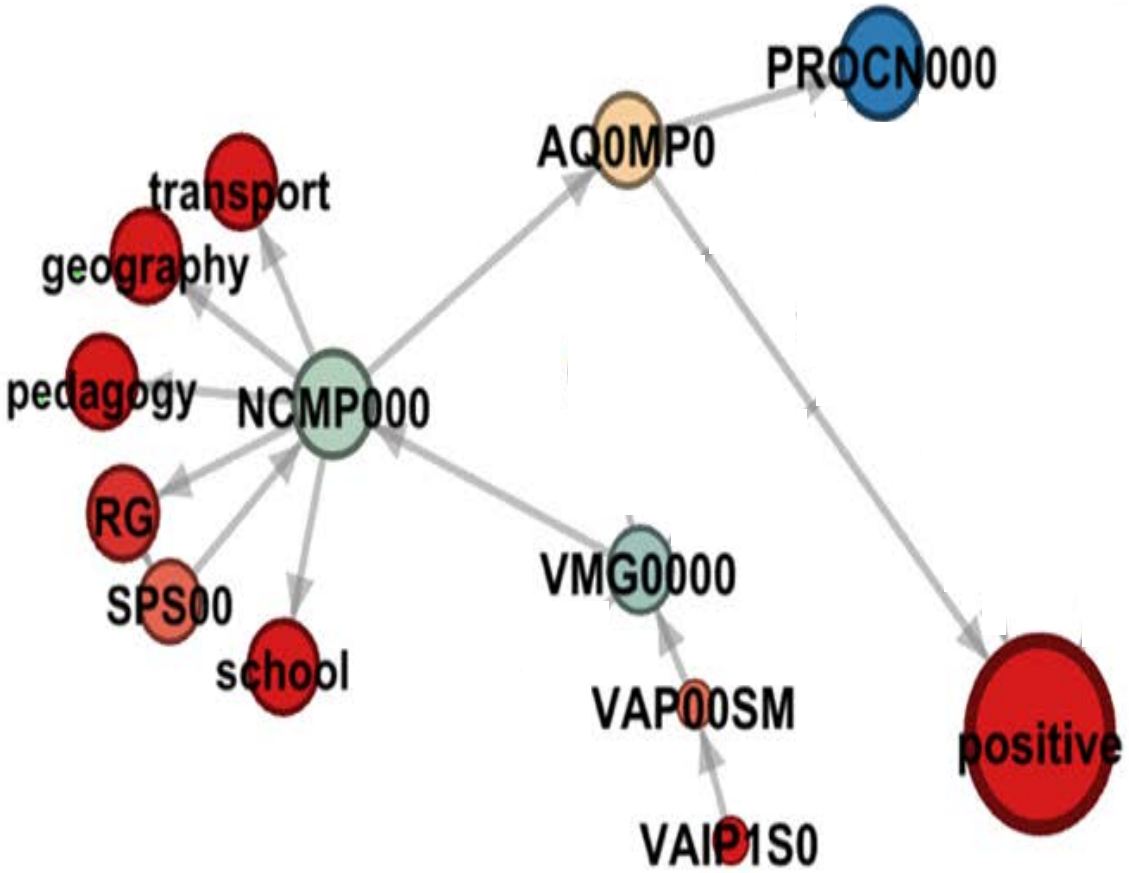
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



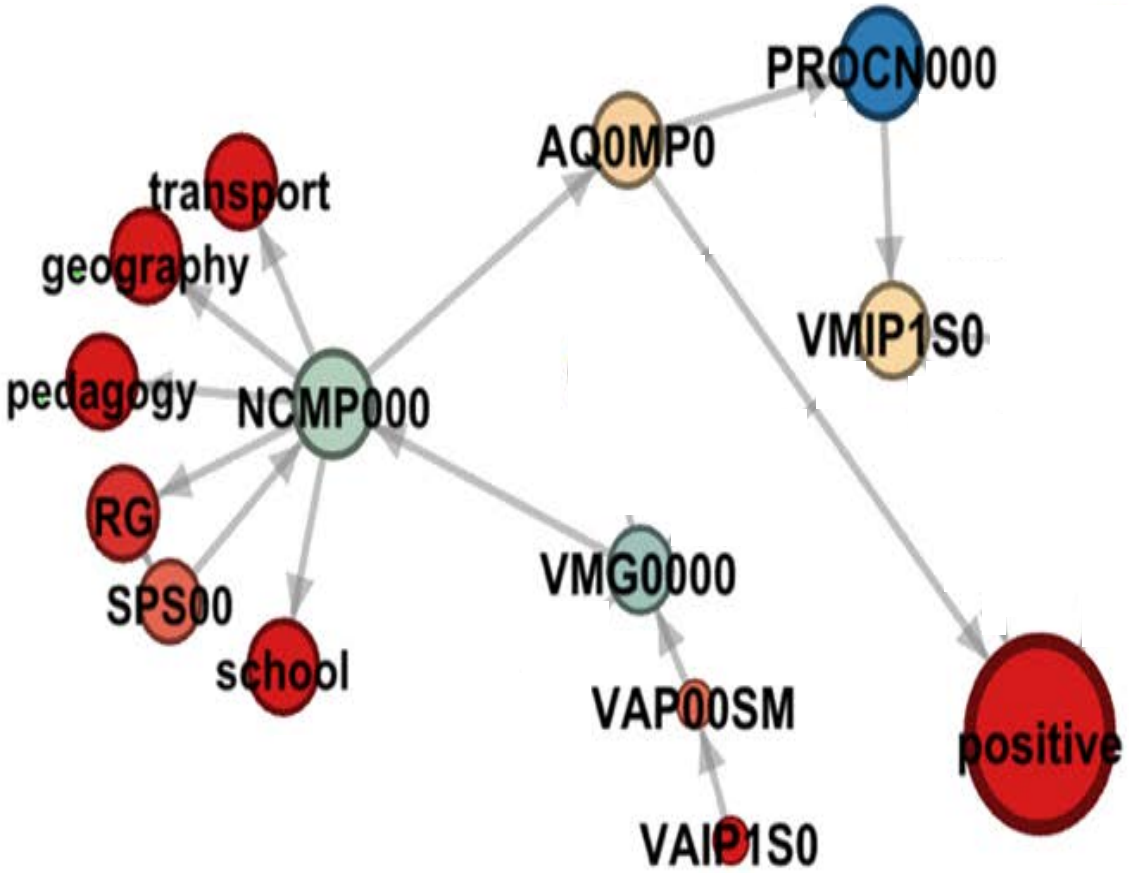
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



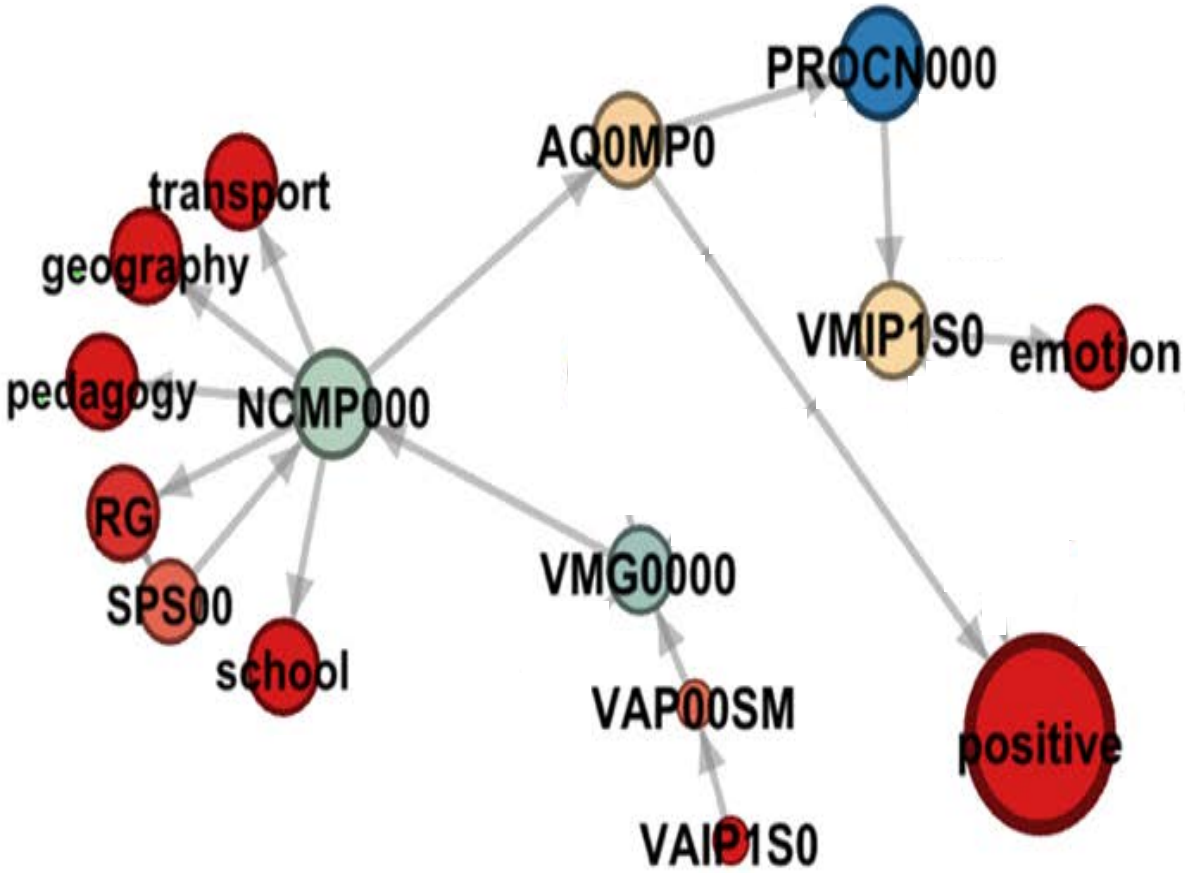
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



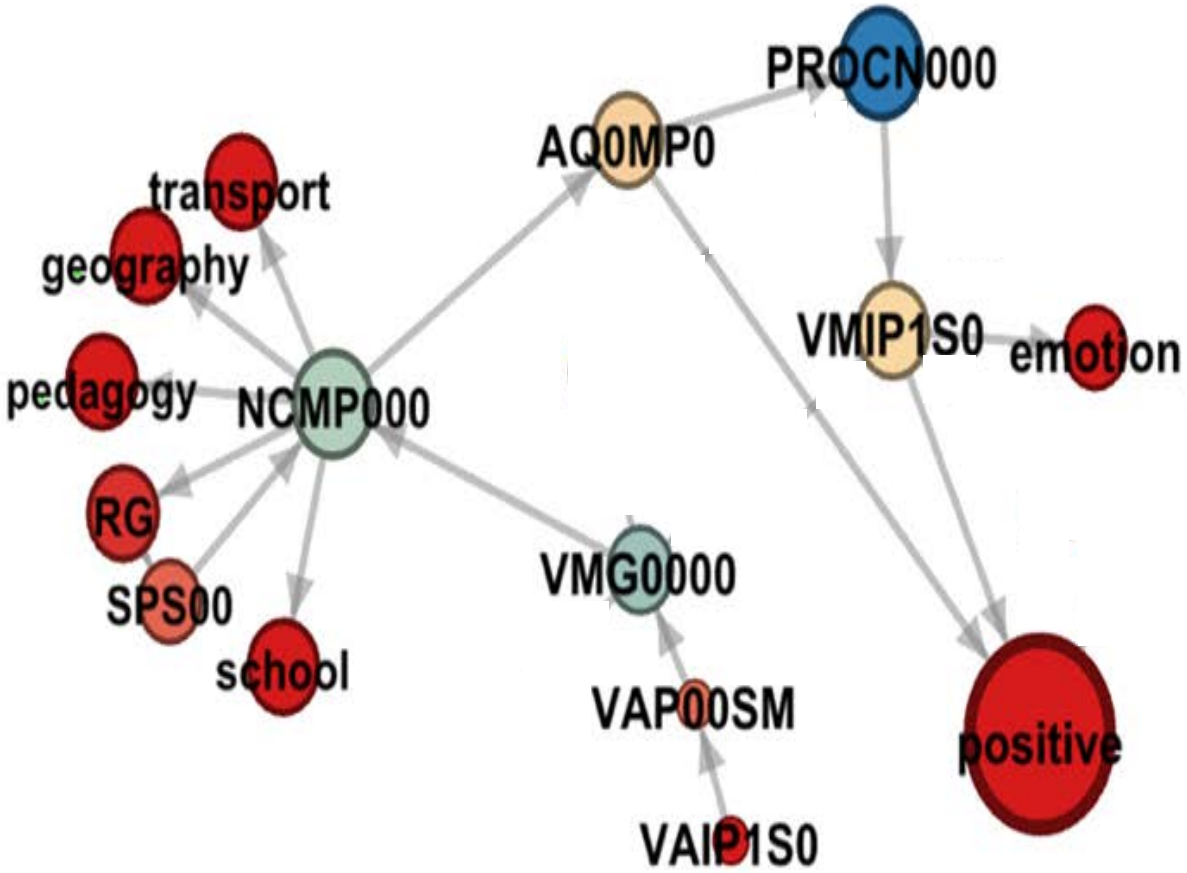
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



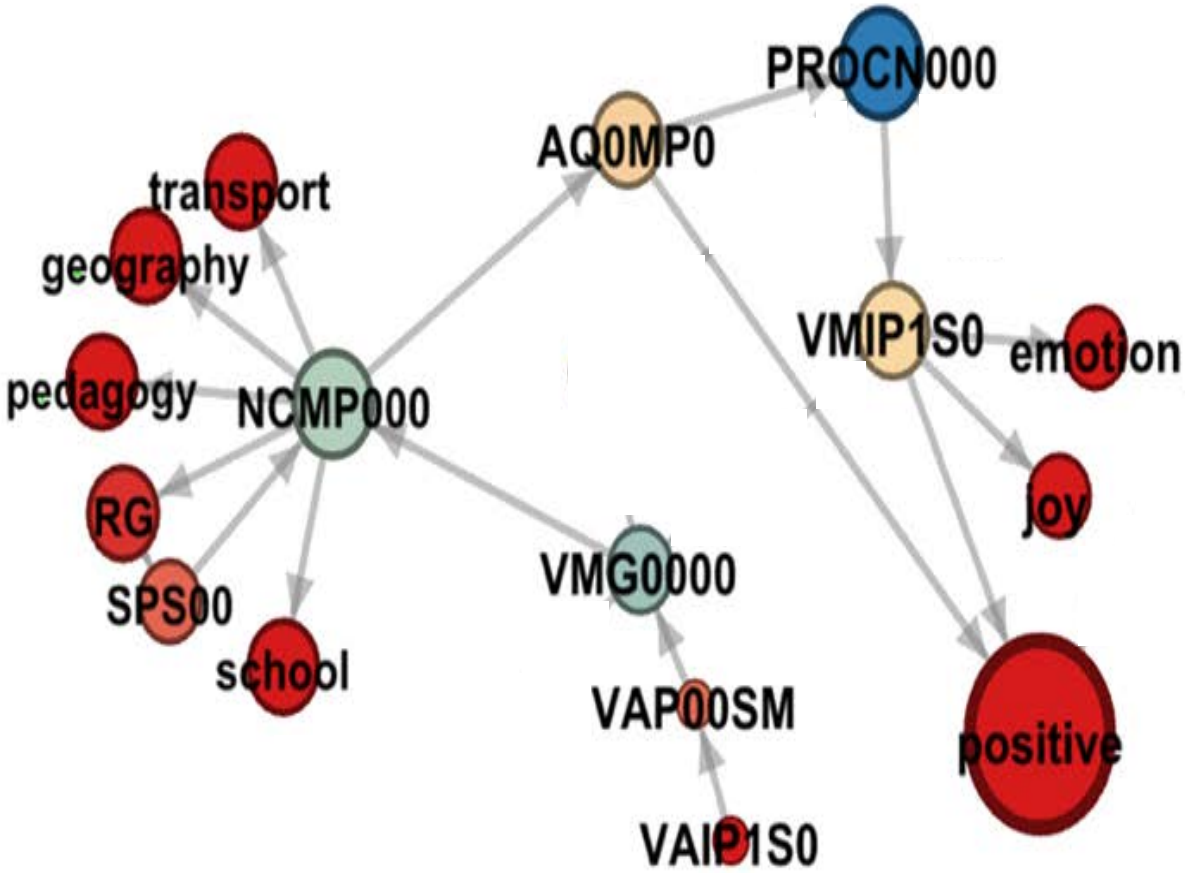
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



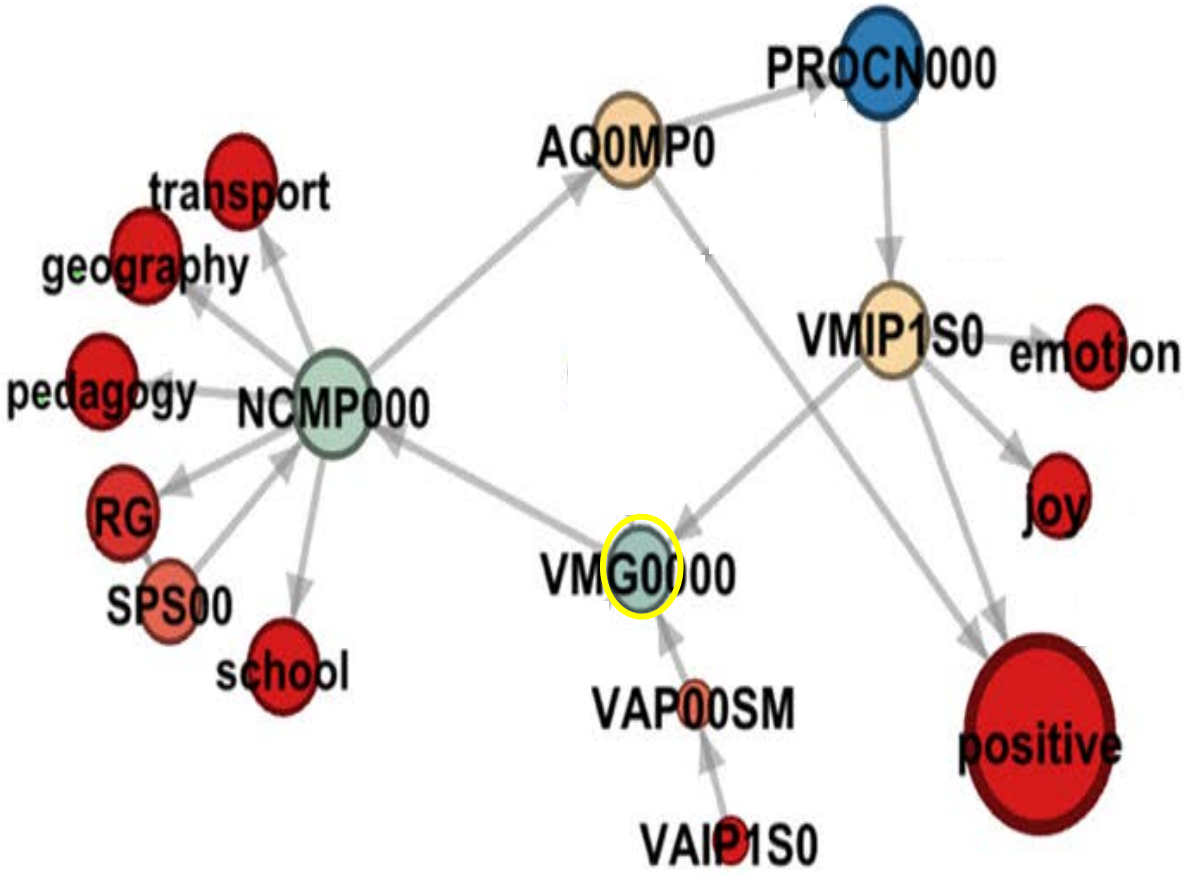
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



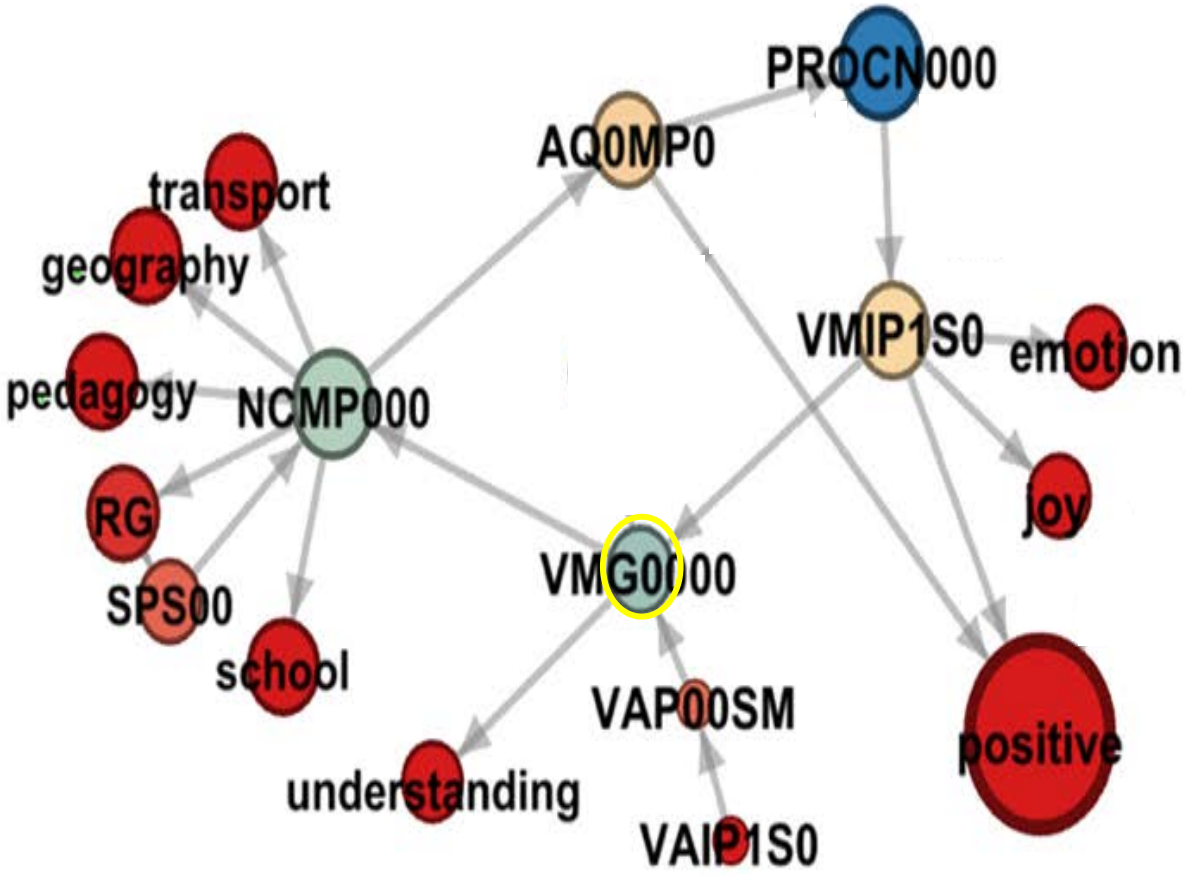
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



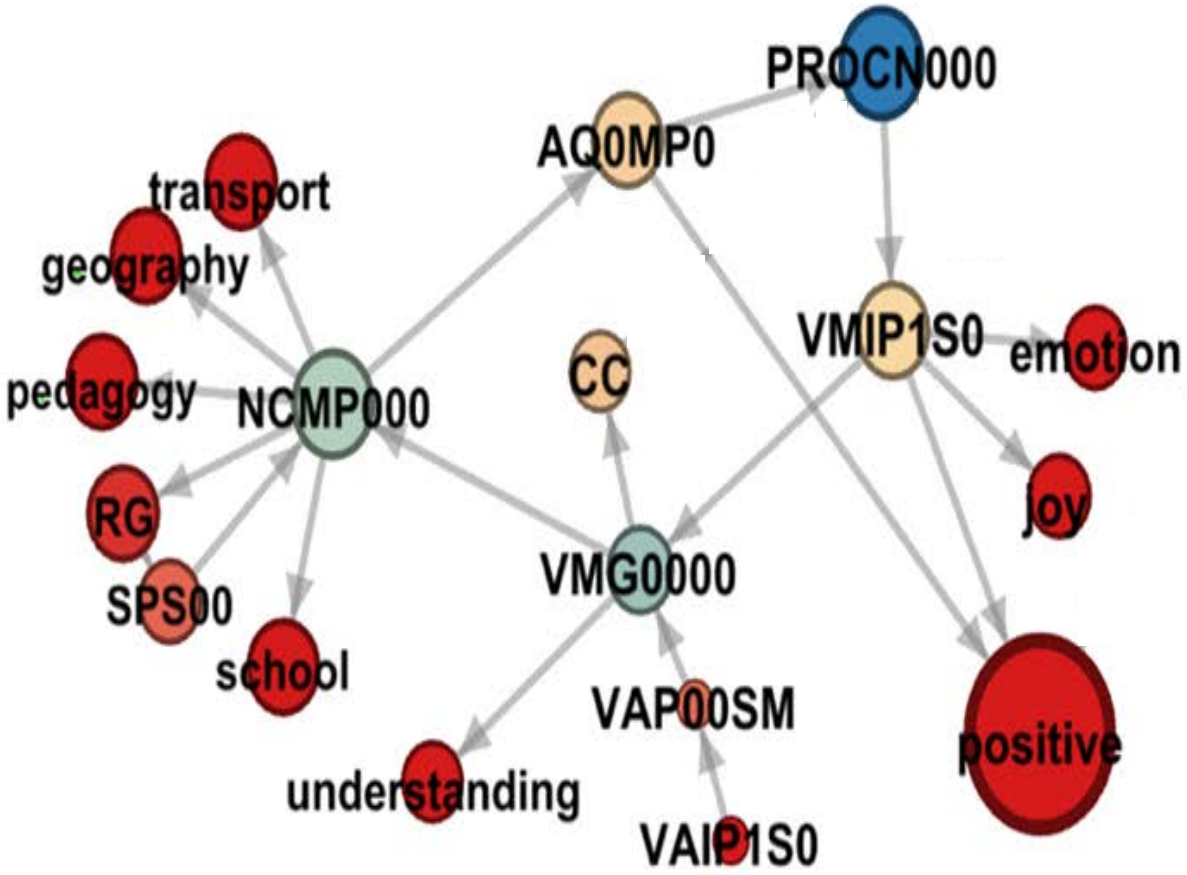
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



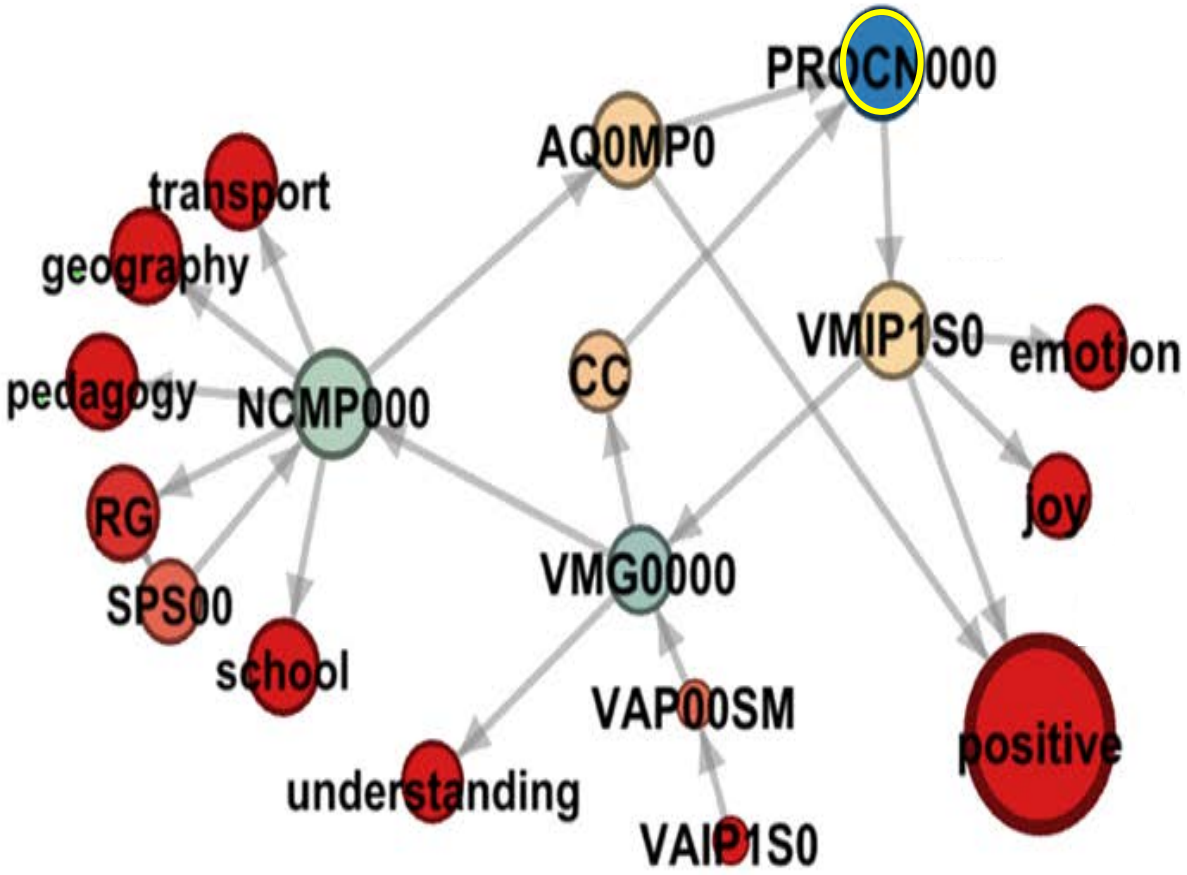
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



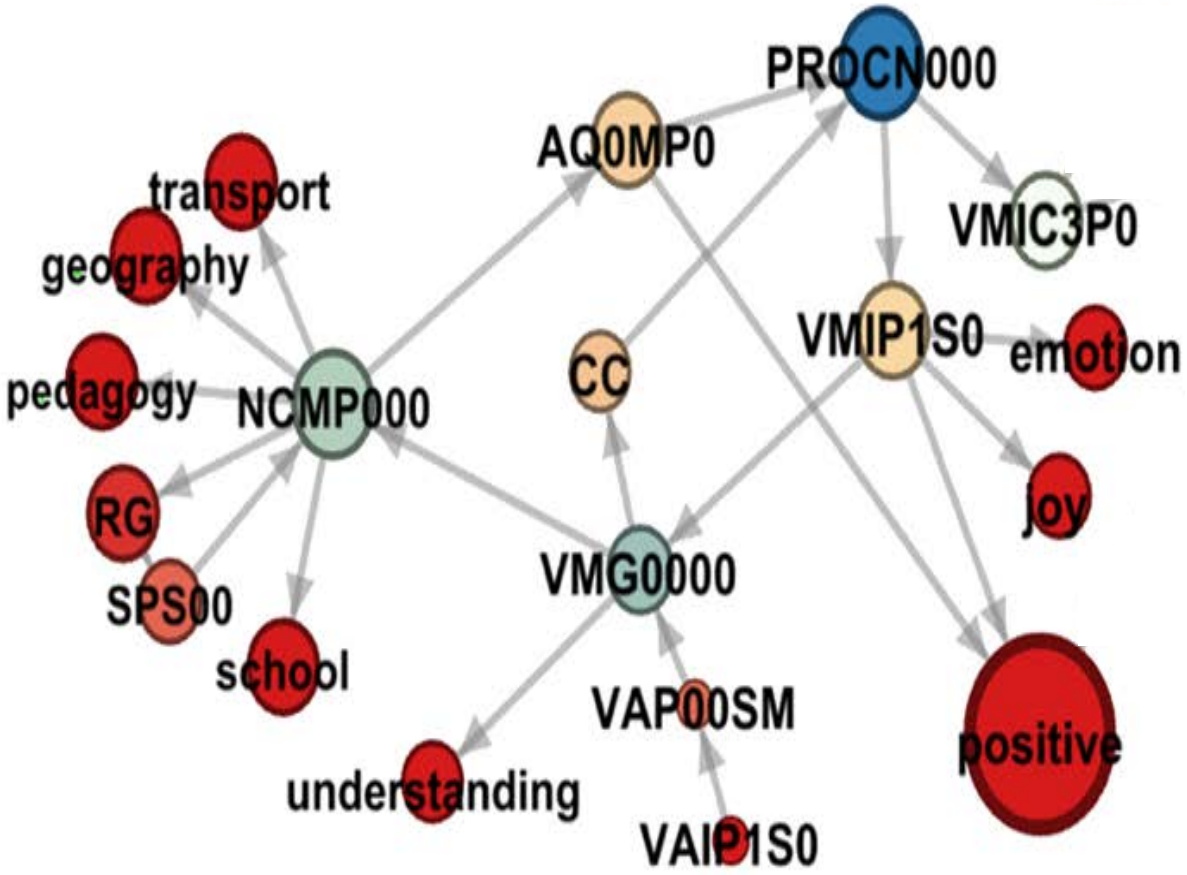
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



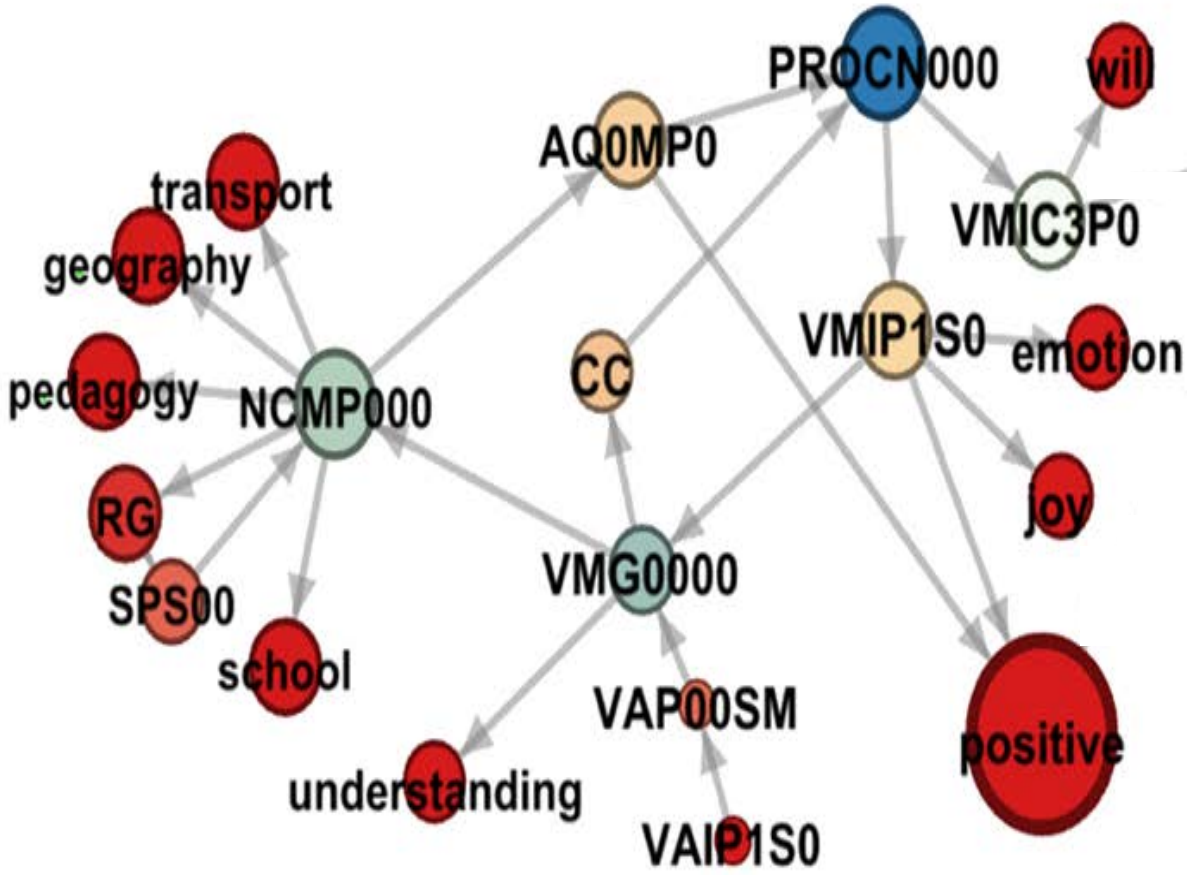
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



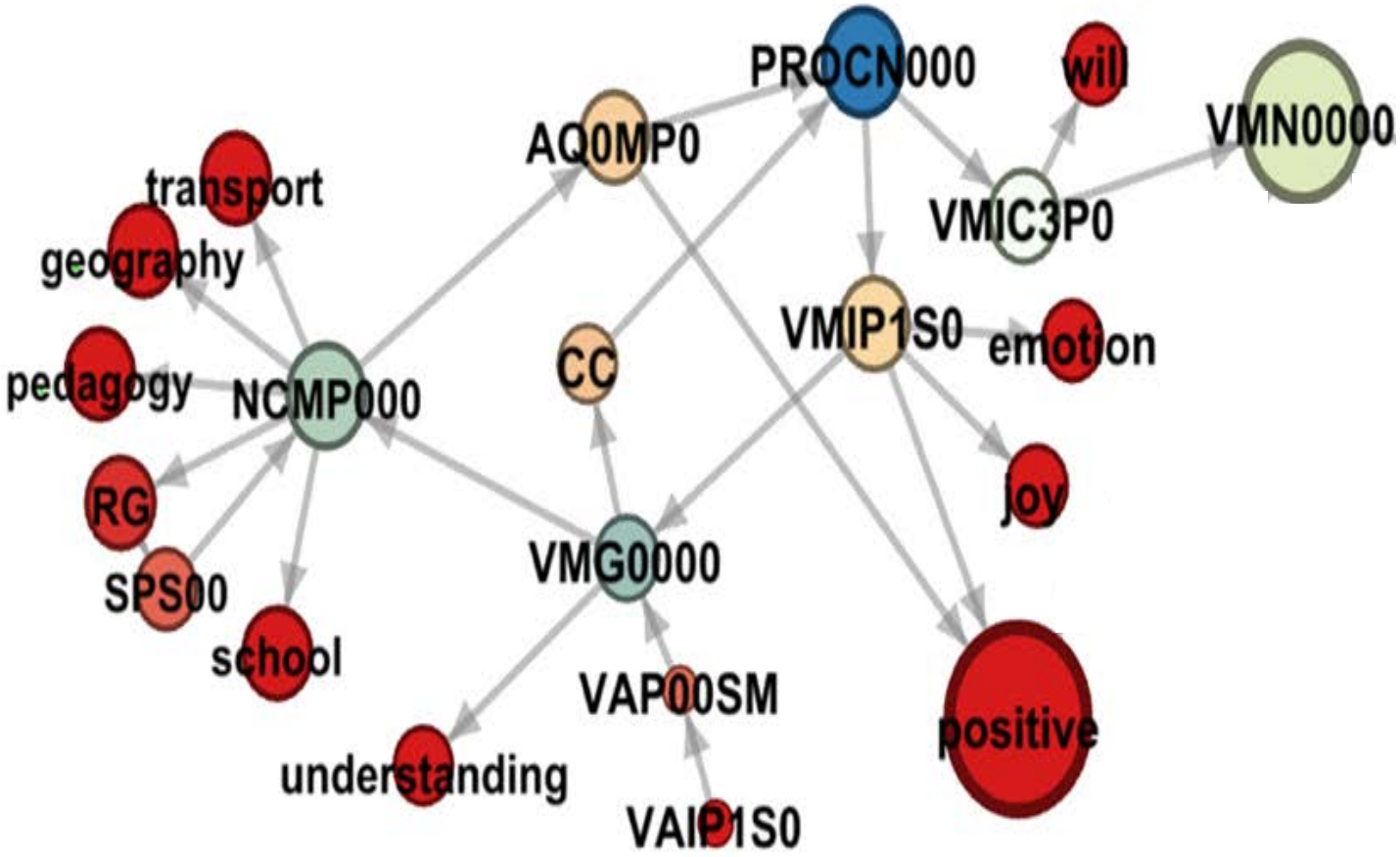
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



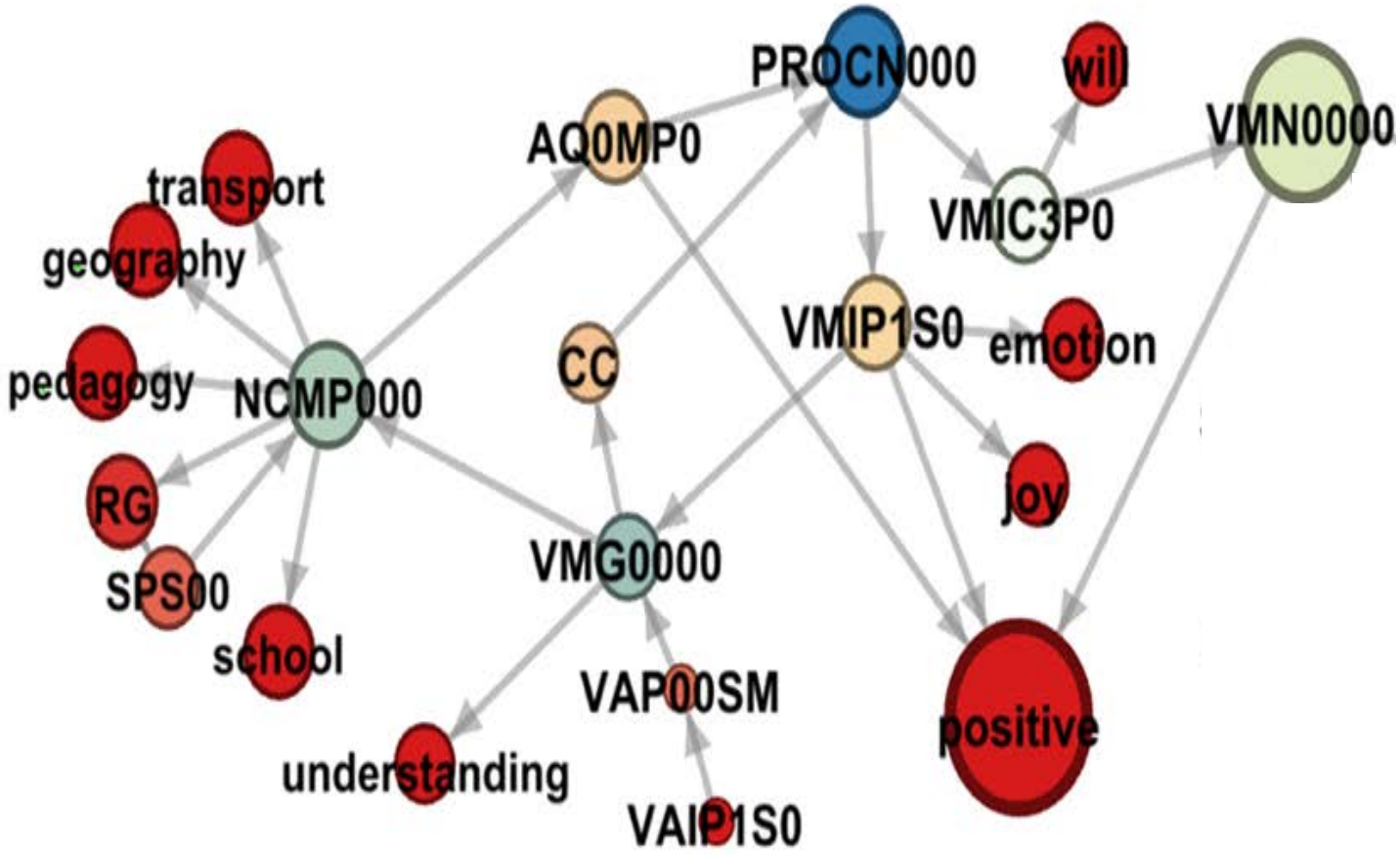
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



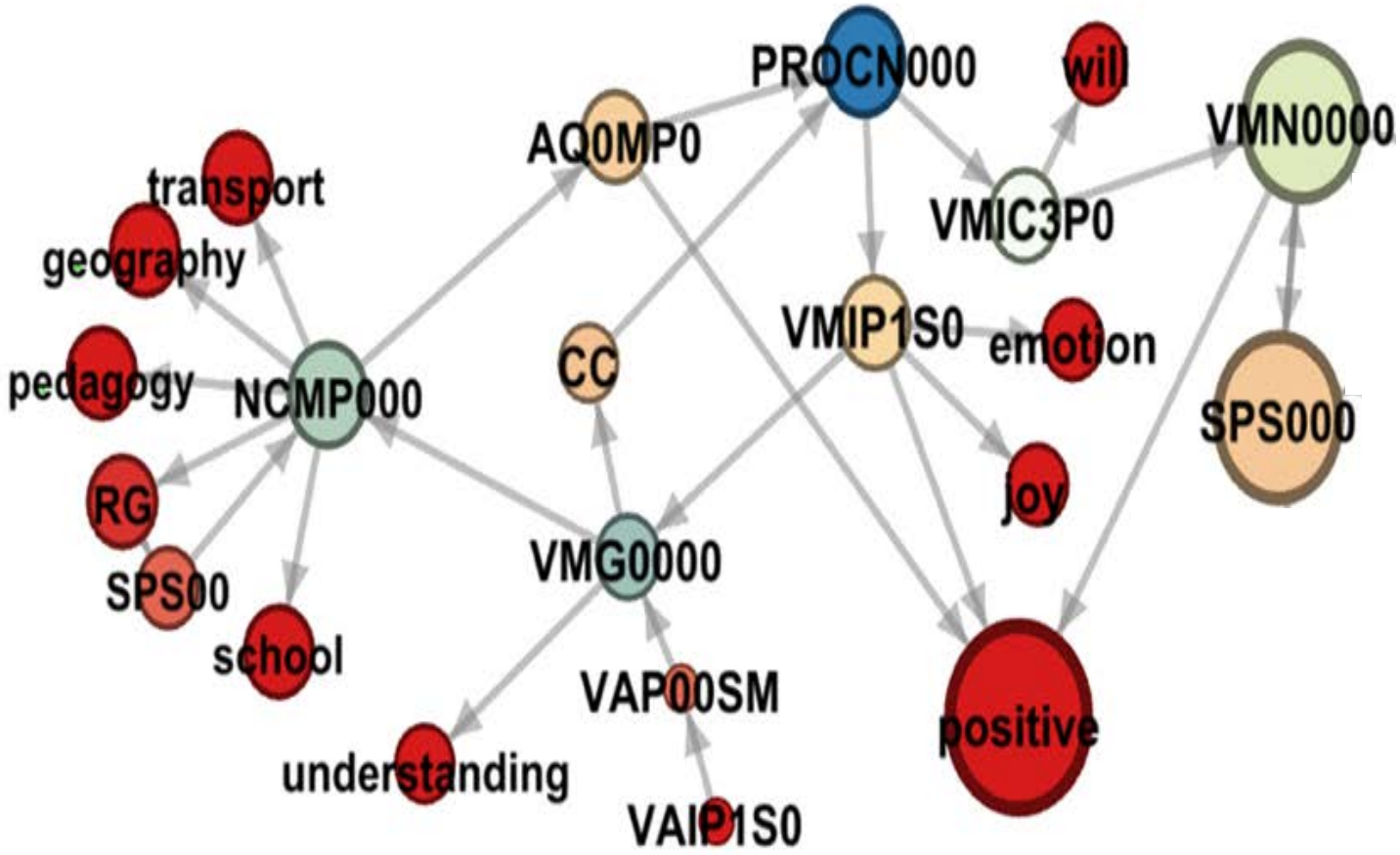
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



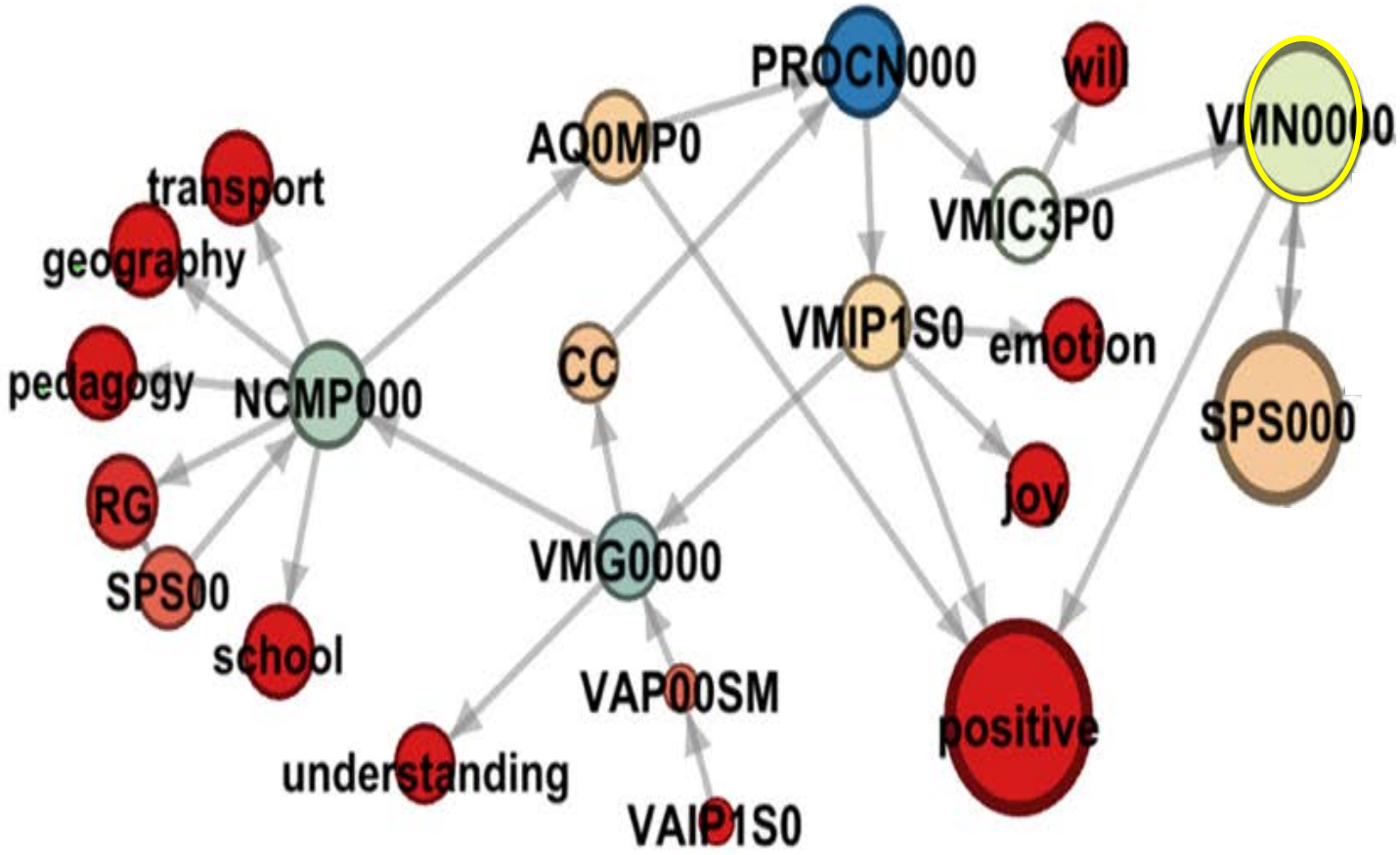
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



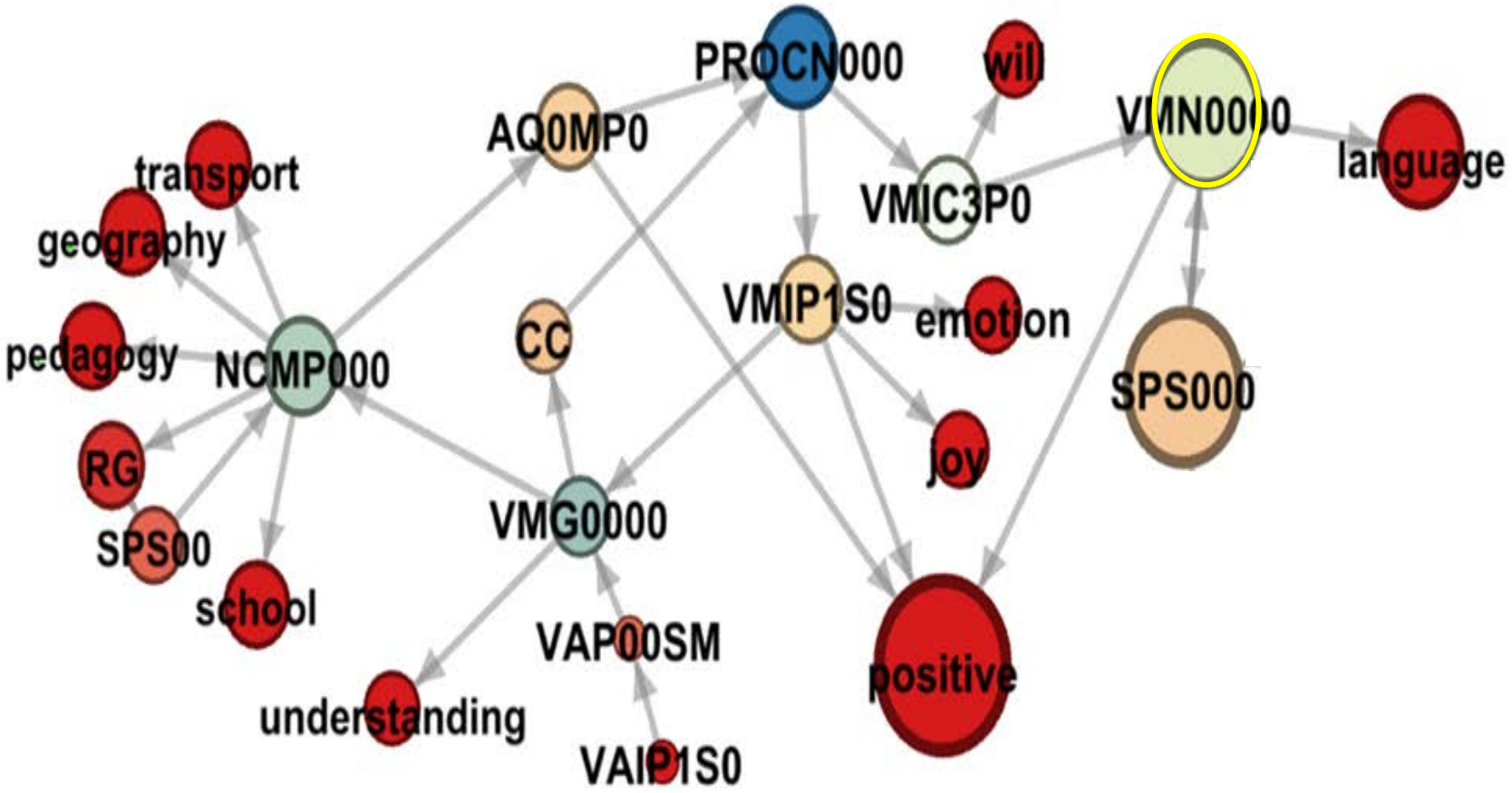
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



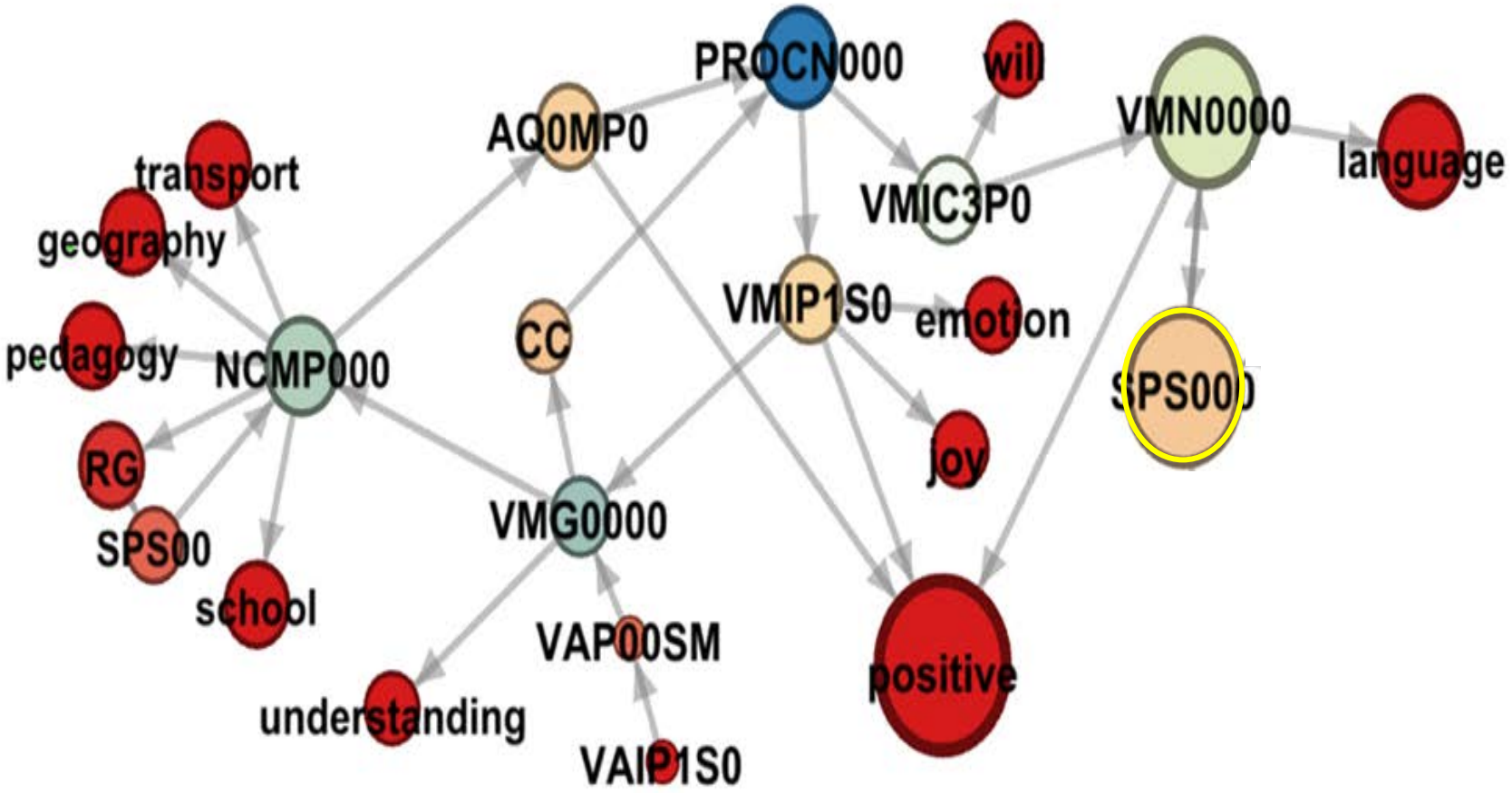
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



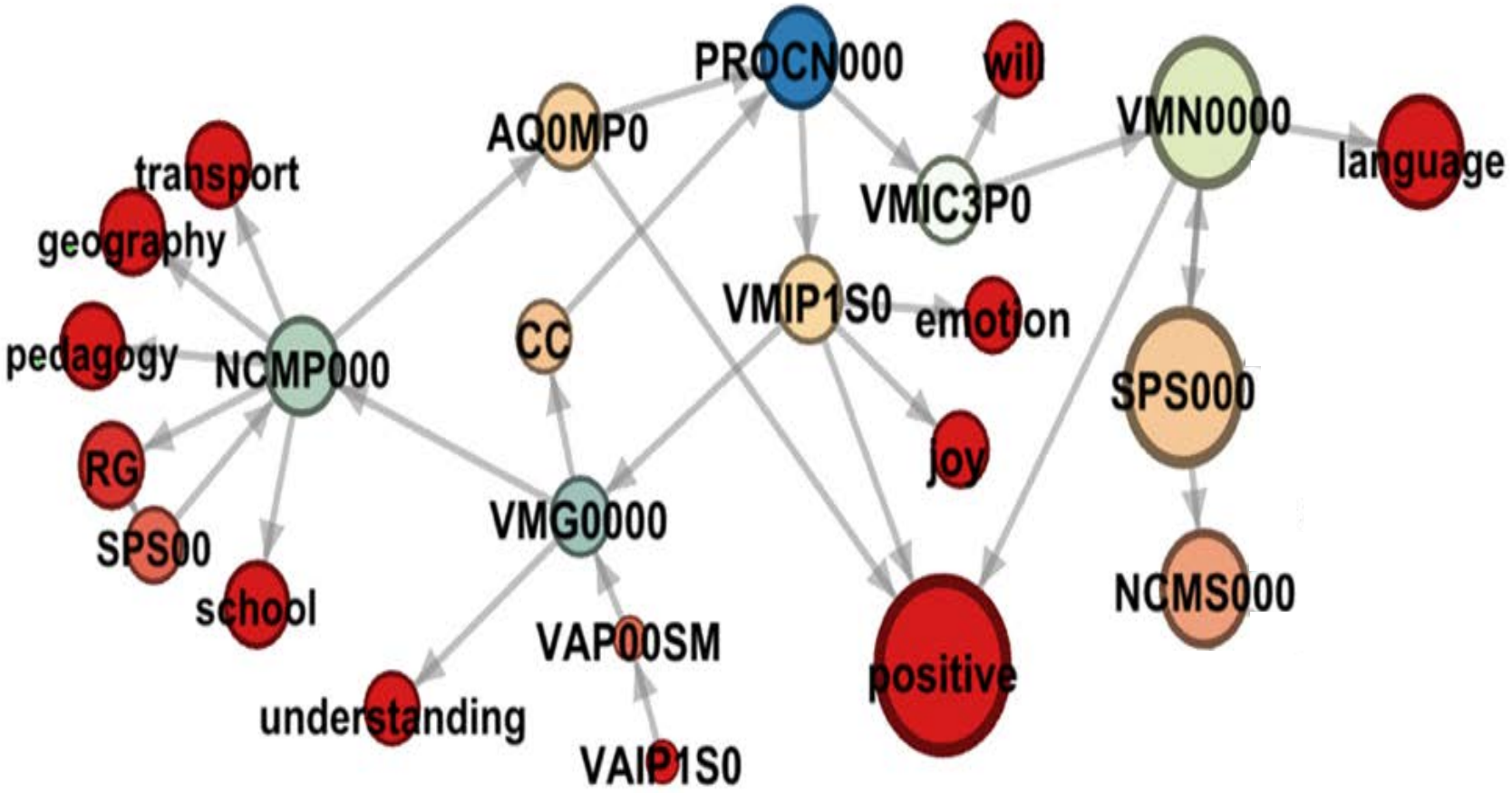
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



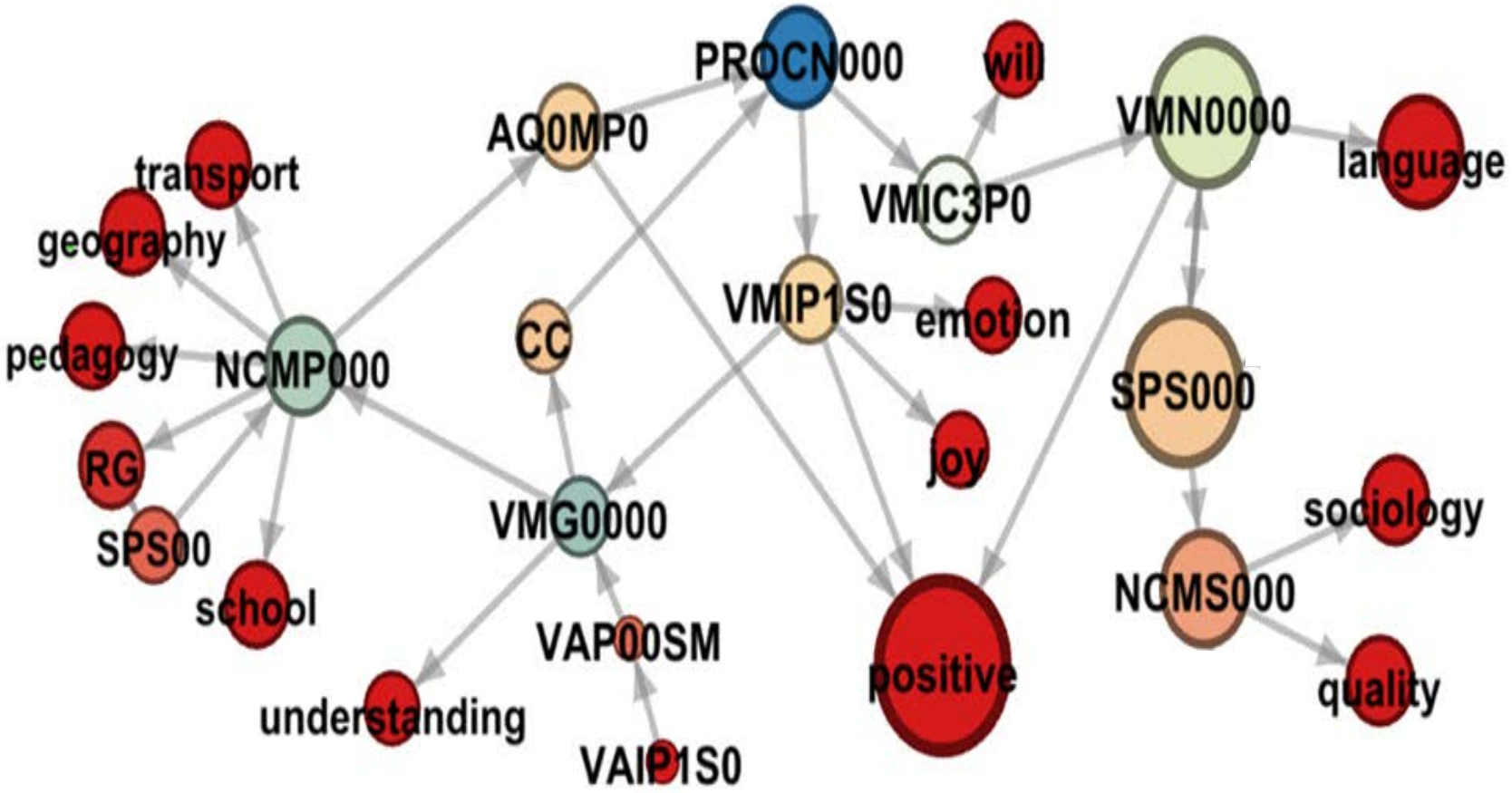
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



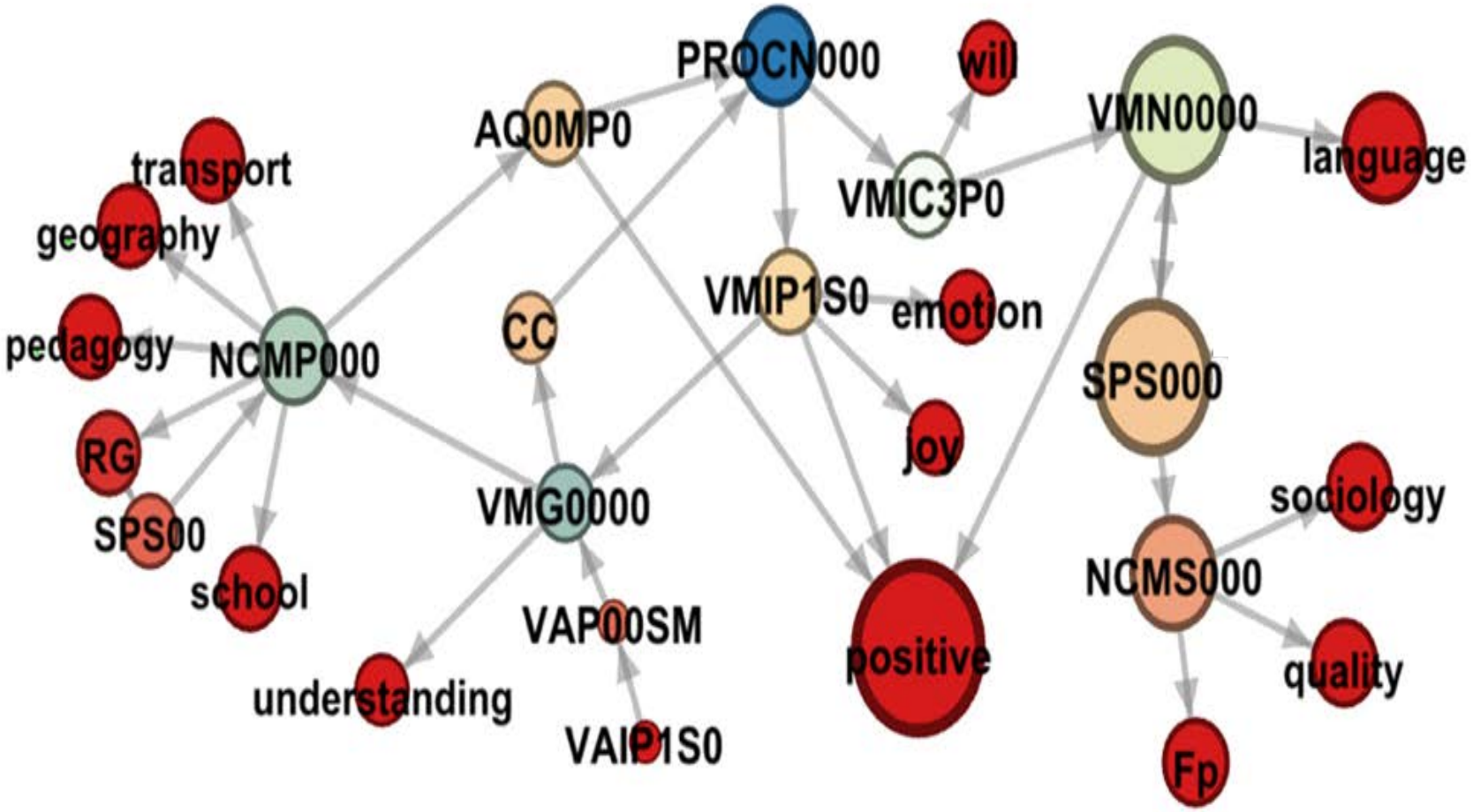
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.

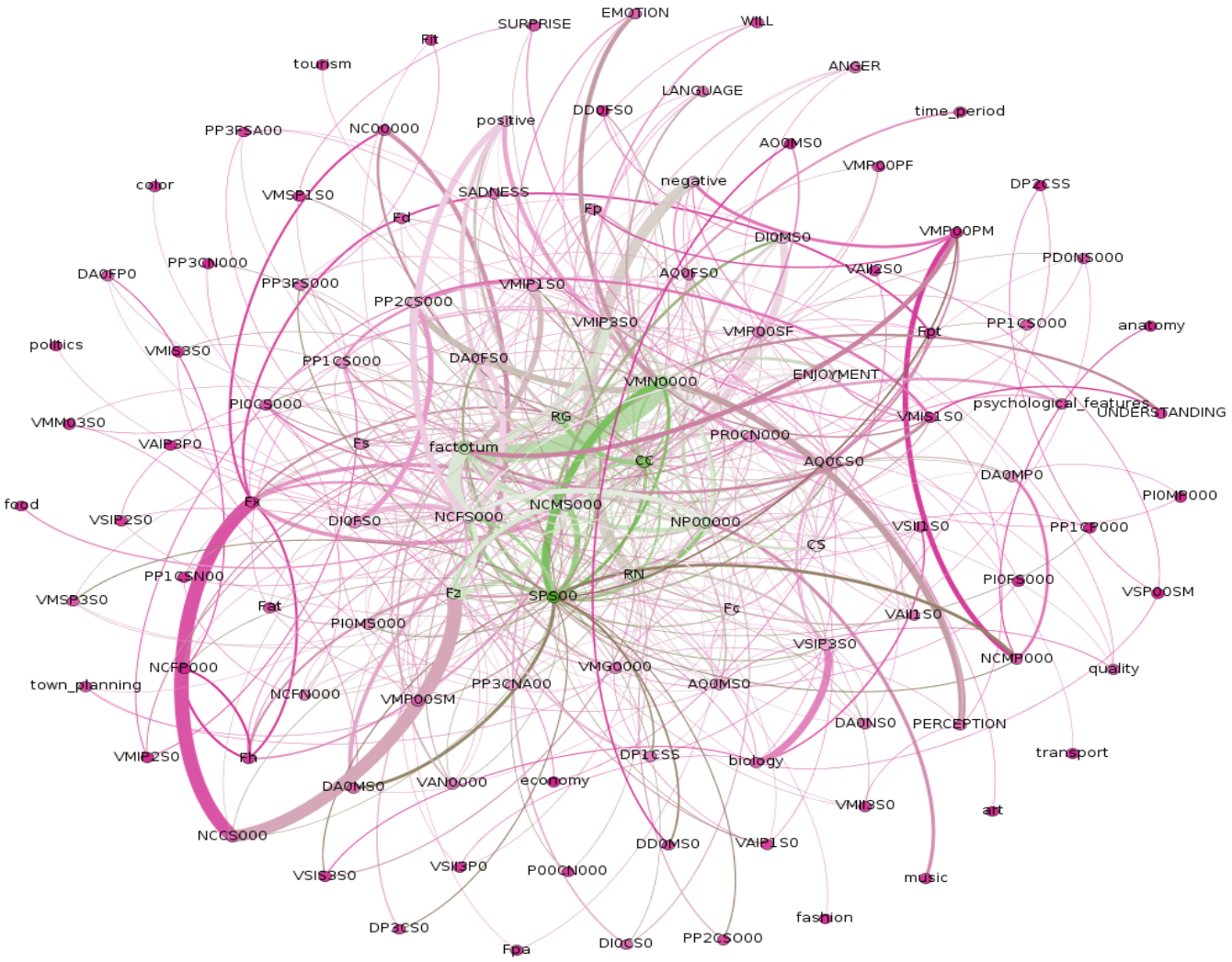


He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



Representation of texts of a class of authors



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

We feed a SVM with…

Structure-based features

Nodes-edges ratio	Indicator of how connected the graph is, i.e., how complicated the discourse is
Weighted average degree	Indicator of how much interconnected the graph is, i.e., how much interconnected the grammatical categories are
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
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 with others
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
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
Average path length	Indicator of how far some nodes are from others, i.e., how far some grammatical categories are from others, or some topics are from some emotions

Node-based features

EigenVector	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, e.g. which nouns, verbs or adjectives
Betweenness	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, e.g. prepositions, conjunctions or even verbs and nouns. Hence, this measure may give us an indicator of what the most common connectors in the linguistic structures used by authors

PAN-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

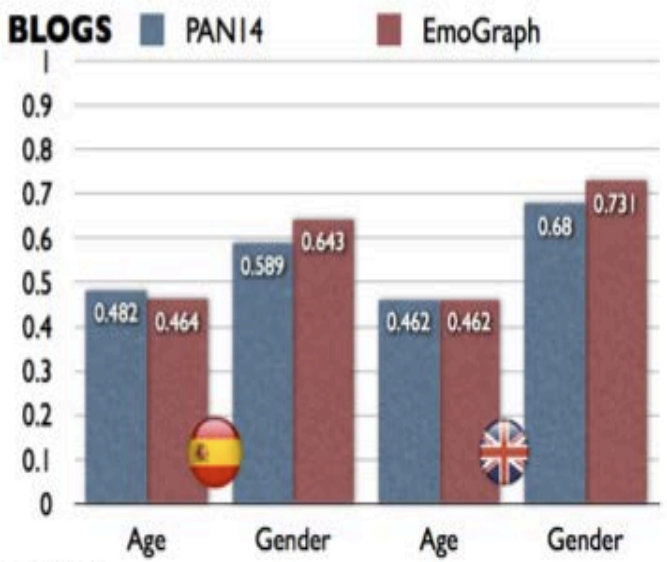
Ranking	Team	Accuracy
1	Santosh	0.6473
2	**Rangel-EG**	0.6365
3	Pastor	0.6299
4	Haro	0.6165
5	Ladra	0.6138
...	...	
8	**Rangel-S**	0.5713
...	...	
18	Baseline	0.5000
...	...	
23	Gillam	0.4784

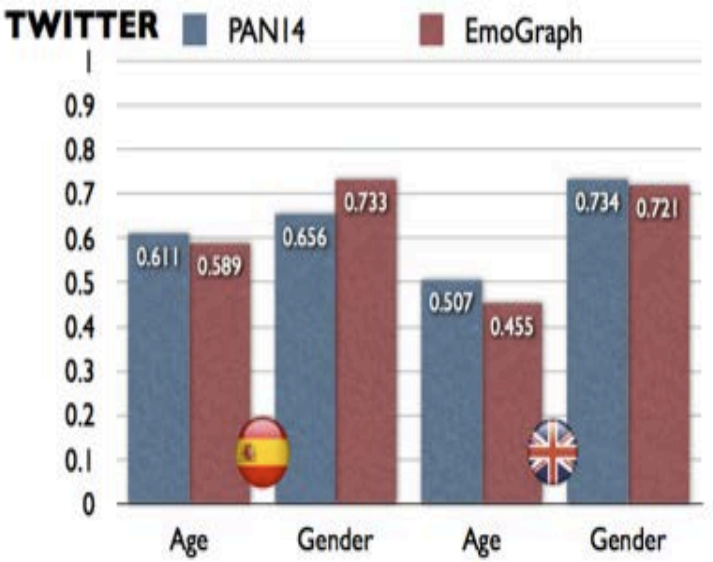
- EG: EmoGraph features based approach
- S: Stylistic features based approach (Ekman's six basic emotions but not in discourse analysis)

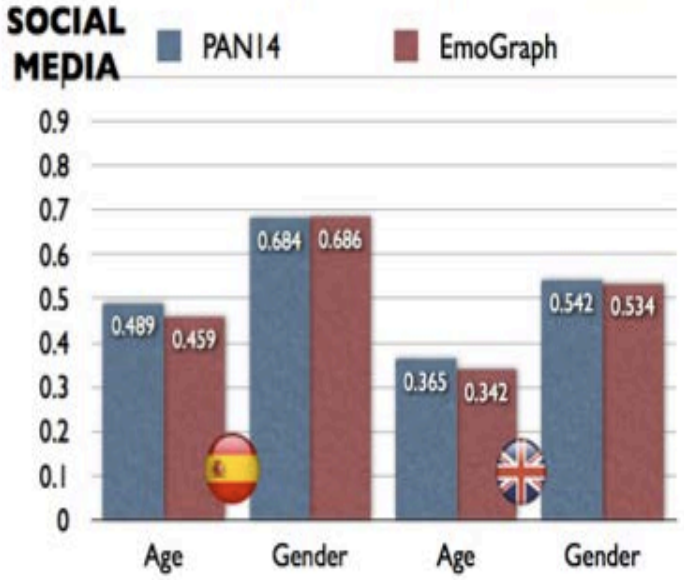
Stylistic + six basic emotions

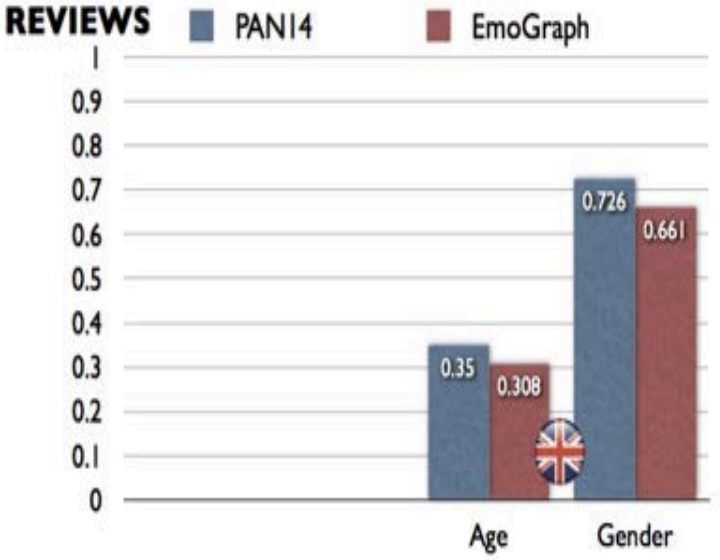
- **Word frequency (F)**: words with character flooding; words starting with capital letter; words in capital letters…
- **Punctuation marks (P)**: frequency of use of dots, commas, colon, semicolon, exclamations and question marks
- **Part-Of-Speech**: frequency of use of each grammatical category
- **Emoticons (E)**: number of different types of emoticons representing emotions
- **Spanish Emotion Lexicon (SEL)**: words co-occurring with each emotion: **happiness, anger, fear, sadness, disgust, surprise**

PAN-14: EmoGraph vs. best approach







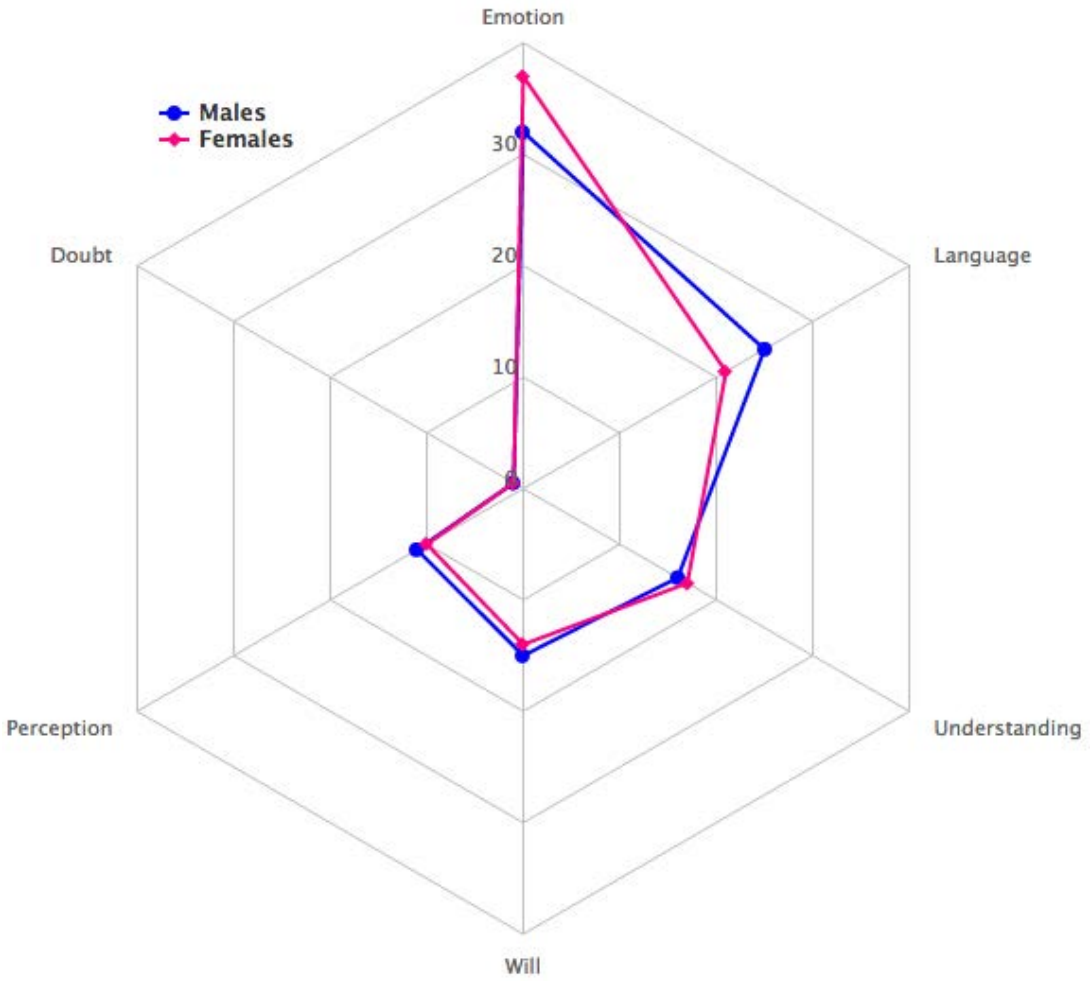


Topics per gender & age (ES)

- Teenagers talk about their studies: e.g. chemistry & linguistics (females) vs. physics & law (males)
- Females talk more about their sexuality and males more about shopping online (commerce)
- We grow up and we are more interested in religion, animals, and food (gastronomy)

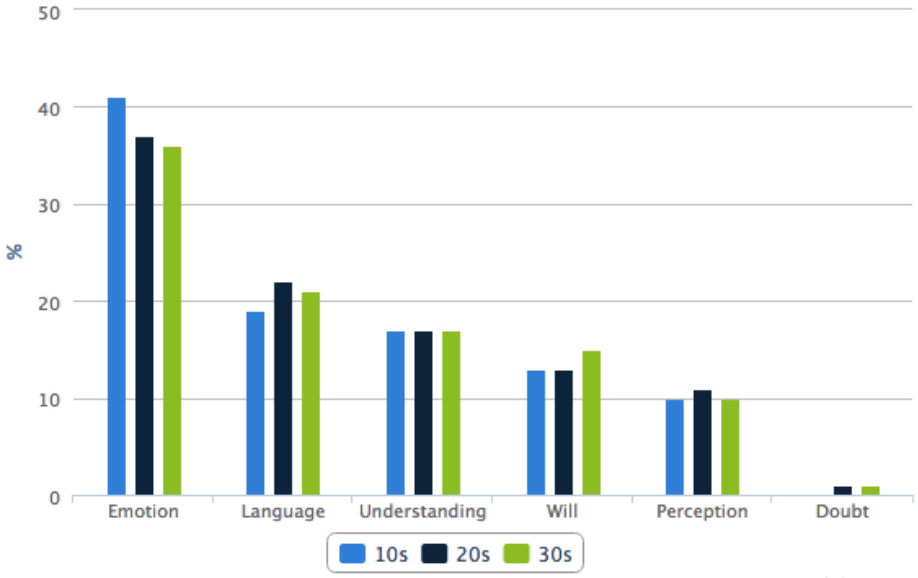
Use of verbs: gender

- **Emotion**: feel, love, want…
- **Language**: say, tell, speak…
- **Understanding**: know, think, understand…
- **Perception**: see, listen…
- **Will**: must, forbid, allow…
- **Doubt**: doubt, ignore…

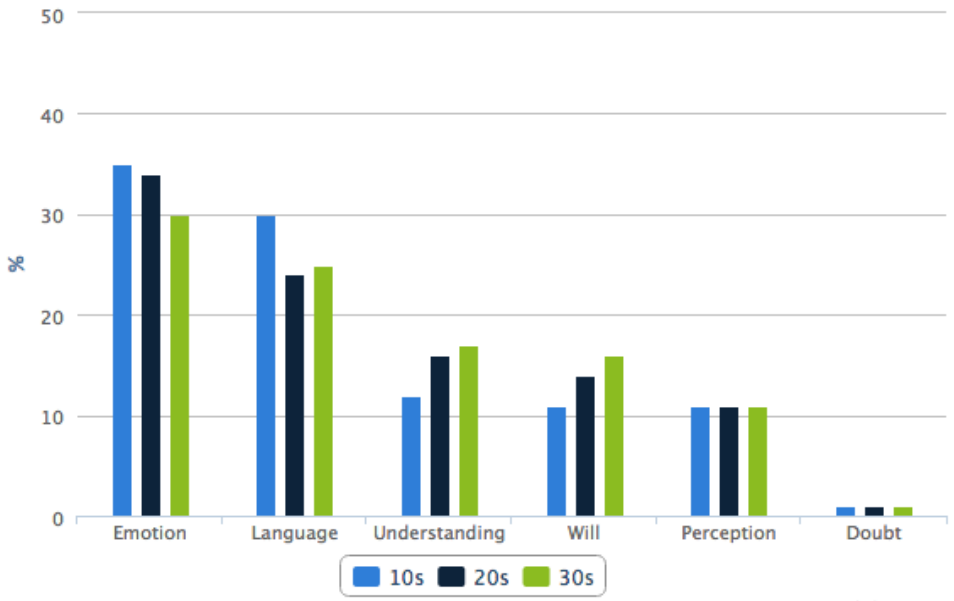


B. Levin. English Verb Classes and Alternations. University of Chicago Press, Chicago, 1993.

Use of verbs: gender & age



Females vs. Males



Profiling native language

Moshe Koppel, Bar-Illan University

…

Native language

Given an English text, can we determine the **author's native language**?

Exercise: which is which?

These were written by **Russian**, **French** and **Spanish** speakers, respectively:

In the second part of this outhor’s novel, called Time Passes, time has passed indeed and Mrs Ramsay has died.

There are pejudments of small groups, such as homosexuals, inmigrants, aids diseaseds, etc. But "political correctness" has have positive and negative consecuences.

There is one more kind of films irritating many television viewers - "soap" serials. «Santa Barbara» has even won "Oskar" prize.

Possible clues

Patterns of native language are typically reflected in how other languages are spoken (Rado, 61, Corder, 81):

- **Word selection**
- **Syntax**
- **Spelling**

Measurable features

- **Frequency of function words**
- **Frequency of letter/char n-grams**
- **Idiosyncrasies** (mistakes)

We will **gather idiosyncrasies data automatically**

Orthographic idiosyncrasies

- Repeated letter (e.g. *remmit* instead of *remit*)
- Double letter appears once (e.g. *comit* instead of *commit*)
- Letter α instead of β (e.g. *firsd* instead of *first*)
- Letter inversion (e.g. *fisrt* instead of *first*)
- Inserted letter (e.g. *friegnd* instead of *friend*)
- Missing letter (e.g. *frend* instead of *friend*)
- Conflated words (e.g *stucktogether*)

Syntactic idiosyncrasies

- Sentence Fragment
- Run-on Sentence
- Repeated Word
- Missing Word
- Mismatched Singular/Plural
- Mismatched Tense
- *that/which* confusion
- Rare POS pairs (Chodorow-Leacock, 00)

Automatically finding idiosyncrasies

1. Run text through automated spell/grammar checker
2. Compare flagged word to best suggestion
3. Mark error accordingly

e.g. text=*remmit* suggestion=*remit*

mark as “repeated letter”

Features

- 400 function words
- 200 letter sequences
- 185 error types
- 250 rare POS pairs

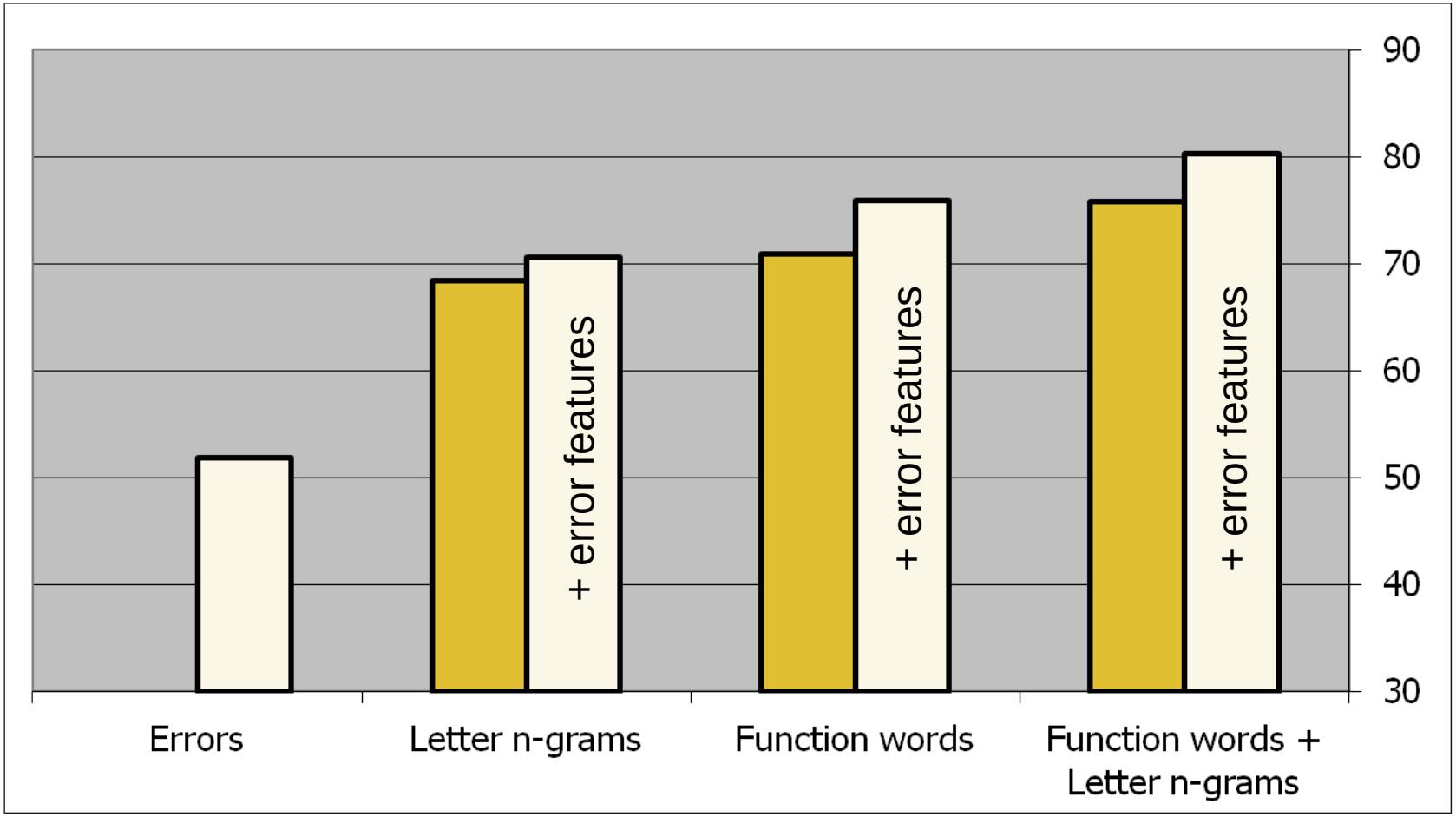
Each document is represented as numerical vector of length 1035

Test corpus

International Corpus of Learner English (Granger, 98)

- 11 countries
- Subjects same age, proficiency level
- Samples same genre, length
- Actually used in study- 258 docs:
  - French
  - Spanish
  - Bulgarian
  - Czech
  - Russian

Classification accuracy (10-fold CV)



Baseline=20% (5 languages)

Some hints

- Russian –*over*, *the* (infrequent), **number_relative-adverb**
- French – *indeed, Mr* (no period), misused *o (*e.g. ***o****uthor)*
- Spanish – *c-q* confusion (e.g. ***c****uality*), *m-n* confusion (e.g. *co****n****fortable*), undoubled consonant (e.g. *co****m****it*)
- Bulgarian – *most*_**adverb**, *cannot* (uncontracted)
- Czech – doubled consonant (e.g. *re****mm****it*)

Exercise: let’s try again…

In the second part of this outhor’s novel, called Time Passes, time has passed indeed and Mrs Ramsay has died.

There are pejudments of small groups, such as homosexuals, inmigrants, aids diseaseds, etc. But "political correctness" has have positive and negative consecuences.

There is one more kind of films irritating many television viewers - "soap" serials. «Santa Barbara» has even won "Oskar" prize.

Exercise: let’s try again…

In the second part of this outhor’s novel, called Time Passes, time has passed indeed and Mrs Ramsay has died.

There are pejudments of small groups, such as homosexuals, inmigrants, aids diseaseds, etc. But "political correctness" has have positive and negative consecuences.

There is one more kind of films irritating many television viewers - "soap" serials. «Santa Barbara» has even won [the] "Oskar" prize.

Profiling sexual offenders

Daria Bogdanova, University City of Dublin
Paolo Rosso, Universitat Politècnica de València
Thamar Solorio, University of Houston

The use of emotions

Six basic emotions [**WordNet-Affect**]

- HAPPINESS (*happy, cheer)*
- ANGER (*annoying*, *furious*)
- FEAR (*scared*, *panic*)
- SADNESS (*bored, sad*)
- DISGUST (*yucky*, *nausea*)
- SURPRISE (*astonished*, *wonder*)

Emotional profile of online sexual predators:

less emotionally stable than mentally healthy people

Bogdanova D., Rosso P., Solorio T. Exploring High-Level Features for Detecting Cyberpedophilia. In: Computer Speech & Language, 28(1): 108-120, 2014

Psychology of the paedophile

- Paedophiles are characterized by **feelings of inferiority, loneliness, low self-esteem and emotional immaturity**

- 60%-80% suffer from other psychiatric illnesses

Paedophilia

- Diagnostic and Statistical Manual of Mental Disorders: *A pedophile is an individual who fantasizes about, is sexually aroused by, or experiences sexual urges toward prepubescent children (generally <13 years) for a period of at least 6 months*

- 88% of child sexual molesters are pedophiles
- 67% of sexual assault victims are underaged
- 19% of children have been sexually approached over the Internet

Sexual offenders in Facebook

Social networks scan for sexual predators with uneven results

Recomendar 1.211 personas han recomendado esto.

Tweet 439

Share

Share this

+1 61

Email

Print

Factbox

Expert advice to kids safe online

http://www.reuters.com/article/2012/07/12/us-usa-internet-predators-idUSBRE86B05G20120712

Sexual offenders in Twitter

Mirror NEWS

Real news, real entert

FRONT PAGE | NEWS | SPORT | 3AM | TV | LIFESTYLE | MONEY

News · UK News · London 2012 Olympics

By Dominic Herbert | 12 Aug 2012 00:27

Twitter paedos exposed: Vile perverts using social networking site to find victims and trade intelligence

Within two minutes of searching we found 20 paedophiles wanting to abuse young children - and 200 in two hours

http://www.mirror.co.uk/news/uk-news/paedophiles-using-twitter-to-find-victims-1253833

http://www.anonews.co/twittergate-twitter/

Perverted Justice

- **Perverted Justice Foundation** investigates and publishes cases of online paedophilia
- **Adult volunteers enter chat rooms as children**. If they are approached they pass information to the police
- The chat data is available at: http://perverted-justice.com
- Some old statistics: **Myspace** (10,786) known sex offenders since 2007); **Facebook**: 2,800 since 2008.

Challenges of automatic detection of online sexual offenders

- Chat data specificity
  - Mistakes, typos, slang (*asl*, *kewl*), character flooding (*hiiiii!*)
- It is easy to provide false information
  - **Paedophiles pretend to be younger** (or of another gender): **fake profile**
  - **Age** (and gender) **prediction** is required

High-level and augmented features

- **High level features** reported to be helpful to detect **neuroticism** level by Argamon et al. (2009)
  - **personal pronouns*** (*I*, *you*)
  - **reflexive pronouns*** (*myself*, *yourself*)
  - **obligation verbs** (*must*, *have to*)
- **Augmented features**
  - **High level features**
  - **Emoticons & Imperative sentences**
  - **Emotional markers**

* J.W. Pennebaker's book: The secret life of pronouns, 2011

Experimental data

1. **Paedophile/Other**
    a. Paedophile/Victim (underaged)
    b. Paedophile/Volunteer
    c. Paedophile/Policeman
2. **Adult/Adult** (consensual relationship)

Experiments

Results of Naive Bayes classification applied to perverted-justice data and the cybersex chat logs:

	Accuracy						
	Augmented features	**High-level features**	**Bag of words**	**Term bigrams**	**Term trigrams**	**Character bigrams**	**Character trigrams**
Run 1	0.93	0.98	0.38	0.55	0.60	0.73	0.78
Run 2	0.95	0.95	0.40	0.50	0.53	0.75	0.45
Run 3	0.95	0.95	0.70	0.45	0.53	0.48	0.50
Run 4	0.98	0.90	0.43	0.53	0.53	0.50	0.38
Run 5	0.90	0.94	0.50	0.48	0.53	0.45	0.50
Average	**0.94**	**0.94**	**0.48**	**0.50**	**0.54**	**0.58**	**0.52**

Results of Naive Bayes classification applied to perverted-justice data and the NPS data:

	Accuracy						
	Augmented features	**High-level features**	**Bag of words**	**Term bigrams**	**Term trigrams**	**Character bigrams**	**Character trigrams**
Run 1	0.93	0.85	0.73	0.60	0.60	0.68	0.75
Run 2	0.95	0.90	0.68	0.53	0.53	0.48	0.45
Run 3	0.95	0.93	0.58	0.53	0.53	0.48	0.85
Run 4	0.98	0.90	0.53	0.53	0.53	0.23	0.80
Run 5	0.90	0.90	0.53	0.53	0.53	0.25	0.75
Average	**0.92**	**0.90**	**0.61**	**0.54**	**0.54**	**0.42**	**0.72**

Profiling irony

Paolo Rosso, Universitat Politècnica de València
Francisco Rangel, Autoritas Consulting
Irazú Hernández, Universitat Politècnica de València

…

Gender and irony in Facebook


Anger


Fear


Disgust


Surprise


Joy


Sadness

Rangel F., Hernández I., Rosso P., Reyes A. Emotions and irony per gender in Facebook. In: Proc. ES³LOD @ LREC-2014, Reykjavík, Iceland, May 26-31, pp. 66-73

Statistics: gender & irony

Annotator	Comments	%
A1	52	4.33
A2	189	15.75
A3	48	4.00

ironic comments per annotator

	Total	%
Ironic	42	3.62
Non-ironic	1158	96.37

ironic/non-ironic comments (2/3 annotators)

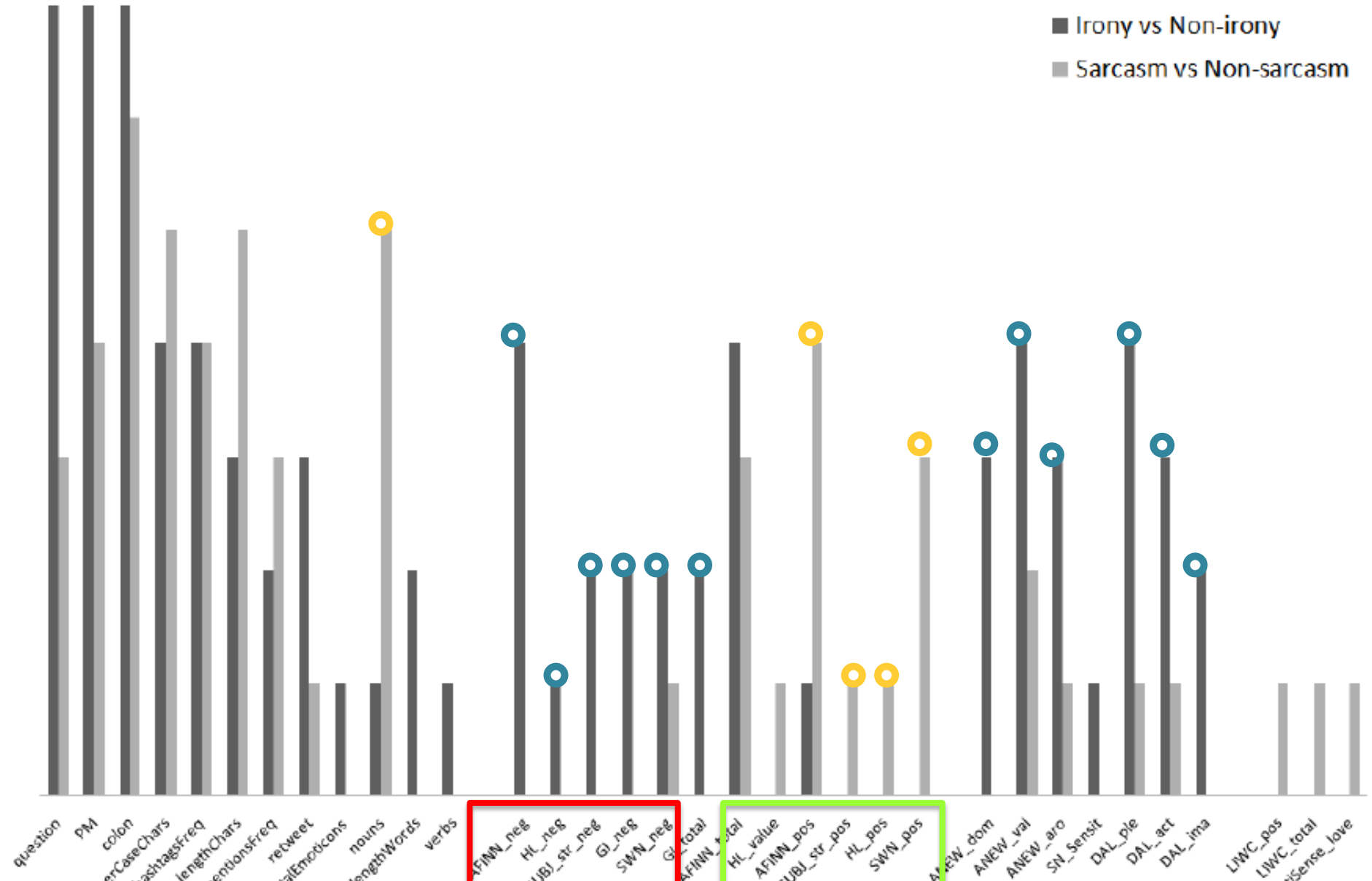
	Female	Male	Total
Football	1	3	4
Politics	11	16	27
Celebrities	3	8	12
Total	15	27	42

ironic comments per topic and gender (2/3 annotators)

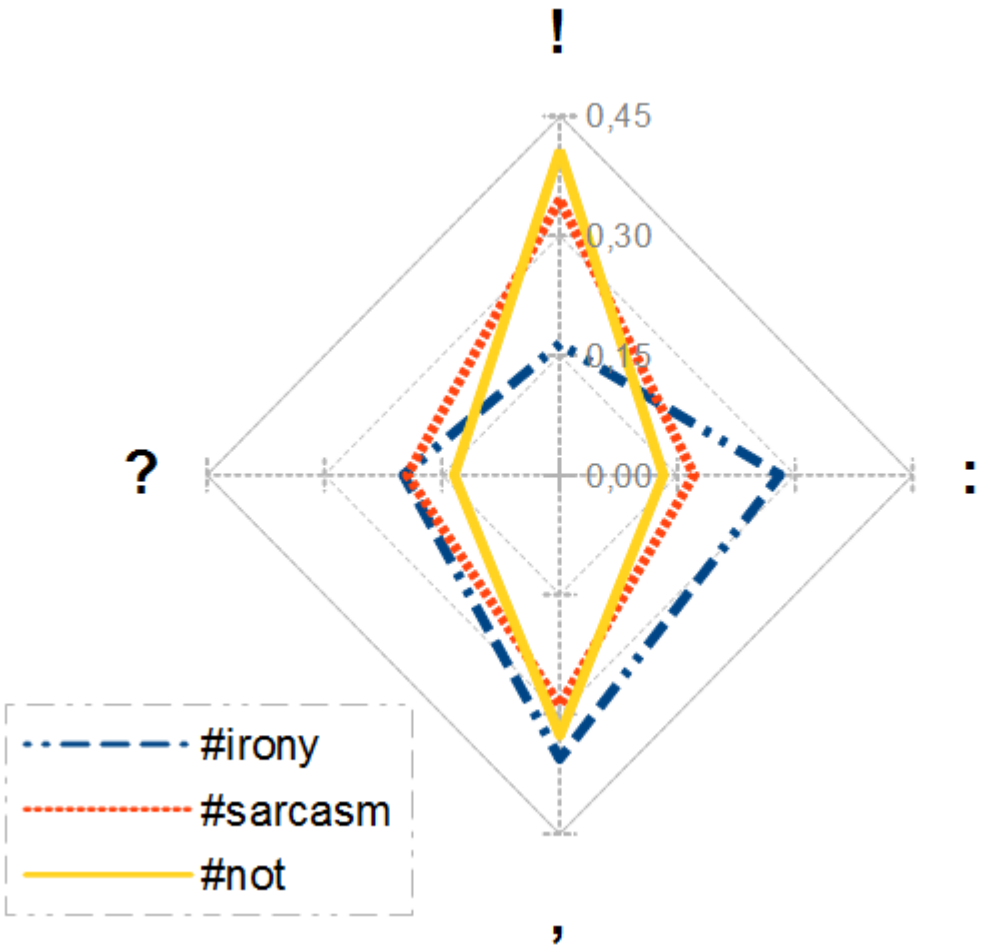
Inter-annotator (dis)agreement

- Fleiss Kappa: It allows multiple annotators (three in our case) and binary variables (ironic / non-ironic)
- We obtained a value of 0.0989: very low index of agreement
  - Irony is quite subjective and depends on annotators, their moods, linguistic and cultural context: we did not provide a common definition for irony
  - Contextual information was not provided, only individual comments
  - Males tended to be more ironic than females (in this corpus)
  - The category politics is the one with more irony and negative emotions and irony: guess why …☹

Sentiment and affective resources

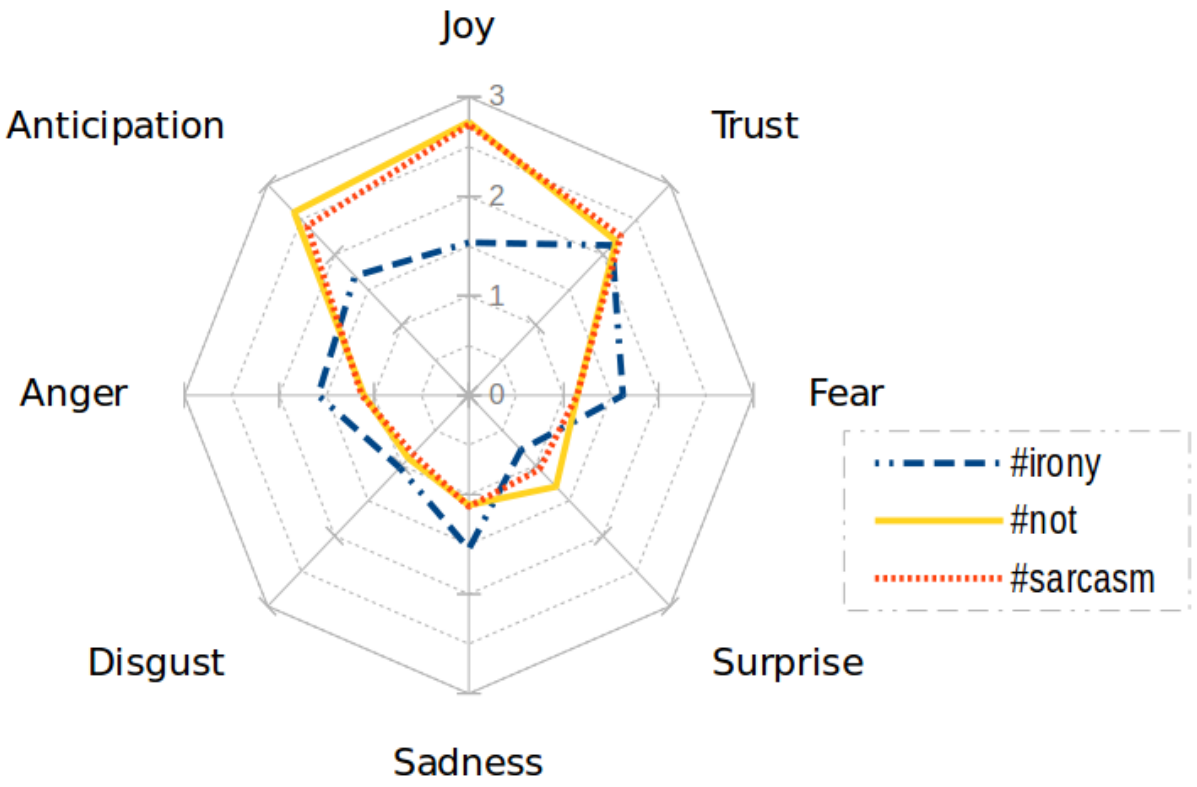


Puntuation marks and length



- ! more frequent for #sarcasm and : for #irony
- Length: #sarcasm tweets are shorter (sarcasm expresses in just few words its negative content)

Distribution of emotions



- #sarcasm: words more related to positive emotions (e.g. Plutchik: joy, anticipation); also in #not
- #irony more creative and subtle: it convey **implicit** emotions (Imagery dimension of Whissel dictionary) whereas #sarcasm more explicit (Dominance dimension of ANEW)

More on irony

Sulis E., Hernández I., Rosso P., Patti V., Ruffo G. **Figurative Messages and Affect in Twitter: Differences Between #irony, #sarcasm, and #not**. In: Knowledge-Based Systems, 108(1): 132-143, 2016

Hernández I., Patti V., Rosso P. **Irony Detection in Twitter: The Role of Affective Content**. In: ACM Transactions on Internet Technology, 16(3): 1-24, 2016

Reyes A., Rosso P. **On the Difficulty of Automatically Detecting Irony: Beyond a Simple Case of Negation**. In: Knowledge and Information Systems, 40(3): 595-614, 2014

Reyes A., Rosso P. Veale T. **A Multidimensional Approach for Detecting Irony in Twitter**. In: Language Resources and Evaluation, 47(1): 239-268, 2013

Thanks
Any question?

Now break 10/15 minutes
After some experiments in R

You can contact me at:
prosso@dsic.upv.es

Don’t forget about PAN @CLEF 2019

Bots and gender profiling

It’s a matter of national pride…

Fratelli d’Italia, l’Italia nel 2019 s’è desta (forse…)