



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Author Profiling en Social Media: Identificación de Edad, Sexo y Variedad del Lenguaje

3 Junio 2016

Autor:
Francisco Manuel Rangel Pardo

Director:
Dr. Paolo Rosso

Inteligencia colectiva

- Grandes cambios socioculturales. Revolución tecnológica.
- Nuevos medios sociales. Nuevos modelos de relación.
- Multitudes inteligentes. Inteligencia colectiva. Big Data.

Problemática

- Anonimato.
- Afán de influencia. Spammers. Usuarios fake.
- Ciberdelincuentes. Acosadores. Pederastas.

Author Profiling

- Dime cómo escribes y te diré quién eres.

Beneficios

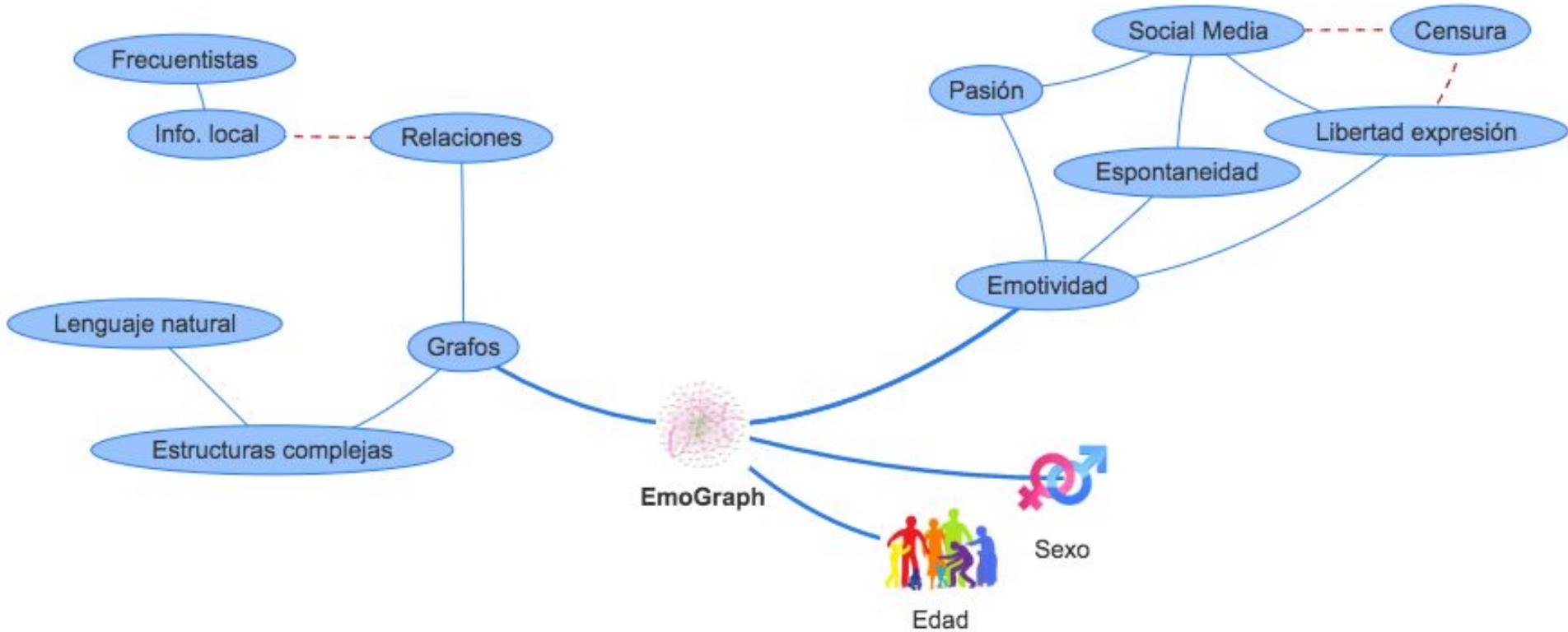
- Mercadotecnia.
- Lingüística forense. Seguridad.



Pennebaker (2003-...)

- Relación entre uso del lenguaje y rasgos de la persona (edad, sexo, ...)
- LIWC: Linguistic Inquiry and Word Count

AUTORES	CORPUS	IDIOMA	APROXIMACIONES
Argamon et al. 2003 Koppel et al. 2003	BNC	inglés	<ul style="list-style-type: none"> • Palabras de función • POS y n-gramas POS • Signos de puntuación • Diccionario • BOW • n-gramas palabras y caracteres • LIWC
Schler et al. 2006 Goswami et al. (2009) Argamon et al. (2009) ...	blogs	inglés	
Peersman et al. (2011) Nguyen et al. (2011-13) Schartz et al. (2013) ...	Netlog Twitter Facebook	holandés inglés inglés	<ul style="list-style-type: none"> • <i>¿y las emociones?</i>



- Hipótesis 1: Relación entre el modo en que expresamos las **emociones** en el marco de nuestro **discurso**, y nuestra **edad** y **sexo**. Además, independientemente de:
 - **Idioma**
 - **Medio social**
- Hipótesis 2: El modo en que expresamos las **emociones** no difiere de la **variedad de la lengua** que hablamos, sino que más bien se difiere en las **palabras/expresiones** que utilizamos para hacerlo.
- Necesidades:
 - ¿Disponemos de los **recursos** adecuados para investigar todas estas cuestiones, incluso en idiomas **diferentes al inglés**?
 - ¿Existe un **marco de evaluación** homogéneo, comparable y reproducible?

EDICIÓN	OBJETIVO	MEDIOS	IDIOMAS	PARTICIPANTES	APROXIMACIONES
2013 (Rangel et al., CLEF 2013)	edad (3 grupos) y sexo	social media	inglés español	21	<ul style="list-style-type: none"> ● Estilísticas ● Emoticonos ● POS ● n-gramas ● LSA ● Diccionario ● IR ● <u>Colocaciones</u> ● <u>2º orden</u> ● ... ● ¿y las emociones?
2014 (Rangel et al., CLEF 2014)	edad (5 grupos) y sexo	social media blogs Twitter revisiones*	inglés español	10	
2015 (Rangel et al., CLEF 2015)	edad (5 grupos), sexo y rasgos de personalidad	Twitter	inglés español holandés italiano	22	

*Sólo en español

Representación preliminar

Conjunto de 59 características basadas en la combinación de:

CARACTERÍSTICAS ESTILÍSTICAS

- Frecuencias (palabras únicas, nº palabras, mayúsculas...)
- Signos de Puntuación
- Categorías Gramaticales con Información Morfosintáctica



EMOCIONES

- Emoticonos
- 6 emociones básicas de Ekman (alegría, sorpresa, disgusto, enfado, tristeza, miedo) (**SEL**) (Sidorov et al., 2012)

Nos referiremos a esta representación como Rangel-S

Experimentación preliminar

Introducción
Sexo y Edad
Variedad del Lenguaje
Conclusiones

Social Media y Author Profiling
Motivación
Cuestiones de Investigación y Objetivos
Laboratorio PAN
Emociones en Social Media

 **Emociones y Tendencias en Twitter**
(Volgmann et al., ECAI 2014)

 **Emociones y Sexo en Facebook**
(Rangel & Rosso, ESSEM 2013)

 **Emociones, Sexo e Ironía en Facebook**
(Rangel et al., LREC 2014)

 **Emociones, Sexo y Edad en PAN**
(Rangel & Rosso, NLPCS 2013)

- Las emociones permite detectar tendencias.
- Estas características permiten identificar:
 - Emociones.
 - Sexo y edad.
- Hemos descubierto* que:
 - Las mujeres usan más emociones.
 - Los hombres son más irónicos.
 - En política se expresan:
 - Más emociones negativas.
 - Más ironía.

*en el dataset EmlroGeFB

Recursos construidos:

- Bárcenas (Twitter)
- EmlroGeFB (Facebook)

- Partiendo de nuestra **hipótesis**:

discurso + emociones --> sexo y edad
- Pretendemos modelar su papel no sólo en base a su frecuencia de aparición, sino por su **posición** con y en **relación** con el resto de elementos del discurso:

eg. *preposición + determinante + nombre + adjetivo*
- **Limitaciones** del estado del arte:
 - Modelos basados en frecuencias no capturan la relación entre elementos.
 - Modelos basados en n -gramas limitados por la elección de la ventana n .
 - Modelos de análisis del discurso fallan en textos altamente informales.

(Rangel & Rosso, IP&M 2016)

Construcción de EmoGraph

Dado un texto:

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

Y los siguientes recursos:

Freeling	http://nlp.lsi.upc.edu/freeling
WordNet Domains + EuroWordNet	http://wndomains.fbk.eu http://www illc.uva.nl/EuroWordNet
Clasificación semántica de verbos	Levin, B. English Verb Classes and Alternations. University of Chicago Press, Chicago. (1993)
Lexicón de polaridad	Hu, M., Liu, B. Mining and Summarizing Customer Reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Seattle, Washington, USA, pp. 168-177 (2004)
Lexicón de emociones	Sidorov, G., Miranda, S., Viveros, F., Gelbukh, A., Castro, N., Velásquez, F., Díaz, I., Suárez, S., Treviño, A., Gordon, J.: Empirical Study of Opinion Mining in Spanish Tweets. 11th Mexican International Conference on Artificial Intelligence, MICAI, pp. 1-14 (2012)

Análisis morfosintáctico con Freeling

He	estado	tomando	cursos	en_línea	sobre	temas	valiosos	que	disfruto	estudiando
VAIP1S0	VAP00SM	VMG0000	NCMP000	RG	SPS00	NCMP000	AQ0MPO	PROCN000	VMIP1S0	VMG0000

y	que	podrían	ayudarme	a	hablar	en	público	.
CC	PROCN000	VMIC3P0	VMN0000	SPS00	VMN0000	SPS00	NCMS000	Fp

Construcción del grafo

He estado tomando cursos en_línea sobre temas valiosos que disfruto estudiando

VAIP1S0 → VAP00SM → VMG0000 → NCMP000 → RG → SPS00 → NCMP000 → AQ0MP0 → PROCN000 → VMIP1S0 → VMG0000 →

y que podrían ayudarme a hablar en público .

→ CC → PROCN000 → VMIC3P0 → VMN0000 → SPS00 → VMN0000 → SPS00 → NCMS000 → Fp

Temas con Wordnet Domains

He estado tomando cursos en_línea sobre temas valiosos que disfruto estudiando

VAIP1S0 → VAP00SM → VMG0000 → NCMP000 → RG → SPS00 → NCMP000 → AQ0MPO → PROCN000 → VMIP1S0 → VMG0000 →

transport
geography
pedagogy
school

y que podrían ayudarme a hablar en público .

→ CC → PROCN000 → VMIC3PO → VMN0000 → SPS00 → VMN0000 → SPS00 → NCMS000 → Fp

sociology
quality

Clasificación semántica de los verbos

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando

VAIP1S0 → VAP00SM → VMG0000 → NCMP000 → RG → SPS00 → NCMP000 → AQ0MPO → PR0CN000 → VMIP1S0 → VMG0000

transport
geography
pedagogy
school

understanding
emotion

y que podrían ayudarme a hablar en público .

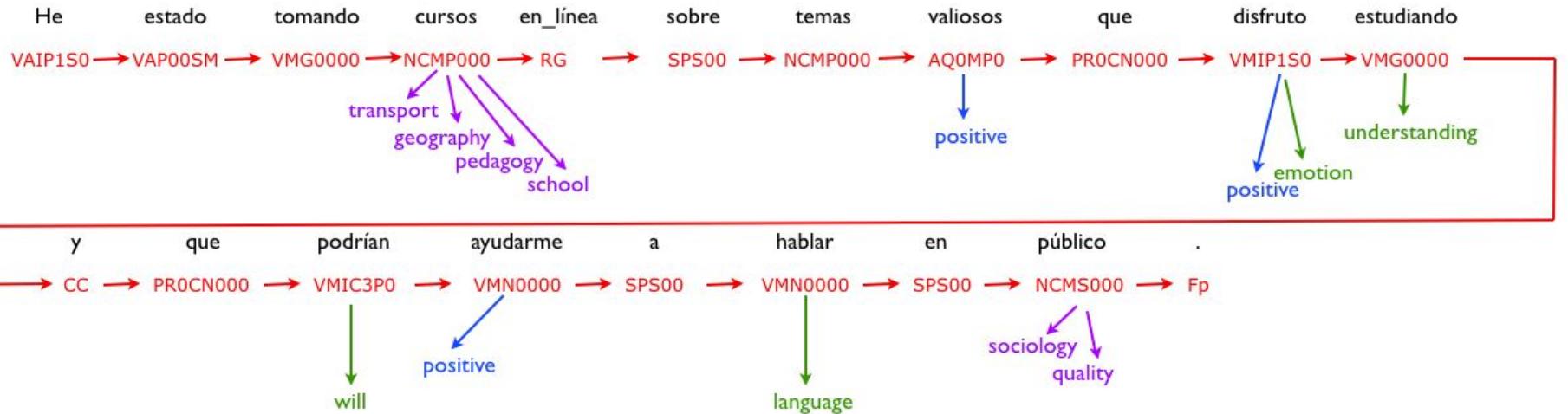
→ CC → PR0CN000 → VMIC3P0 → VMN0000 → SPS00 → VMN0000 → SPS00 → NCMS000 → Fp

will

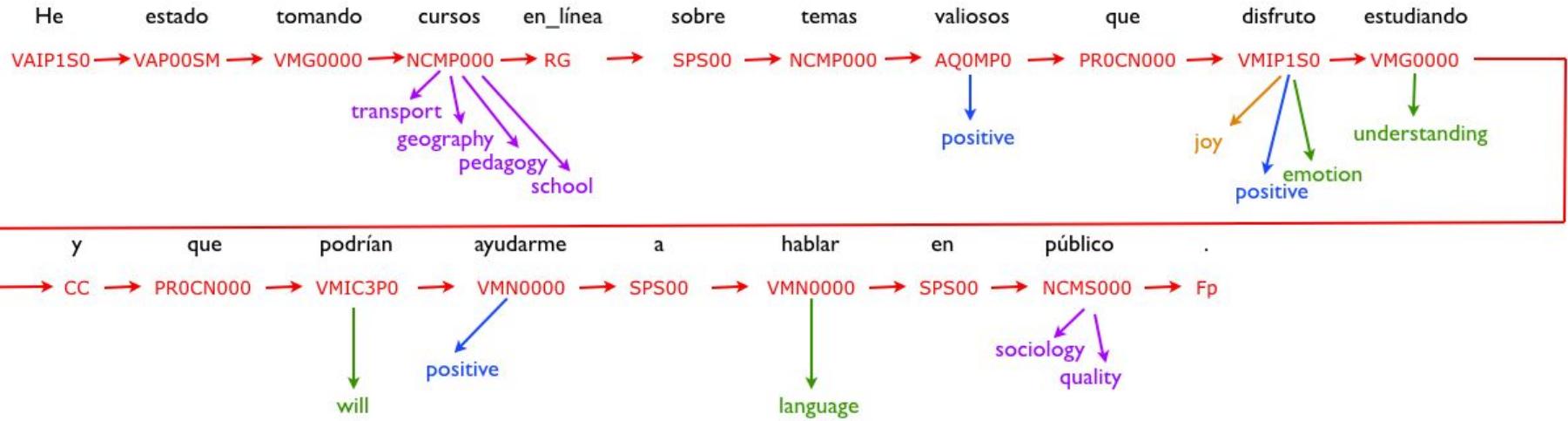
language

sociology
quality

Polaridad



Emociones



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

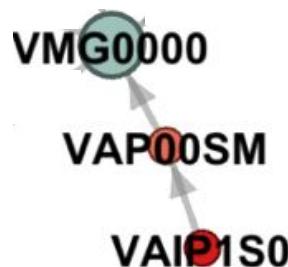


He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

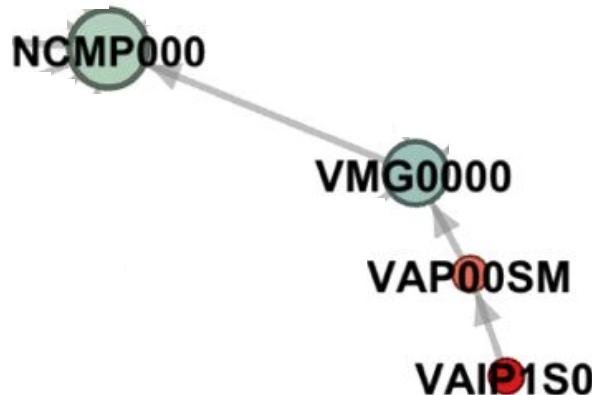


VAP00SM
VAIP1SO

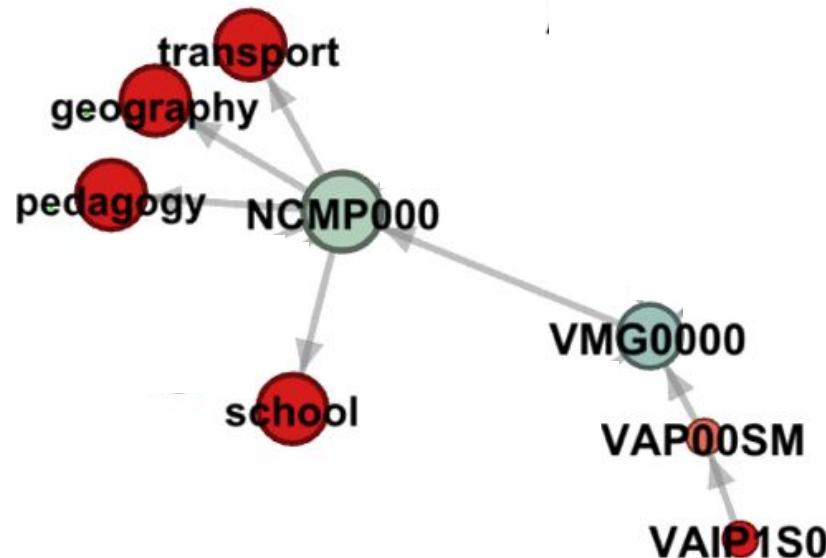
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



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

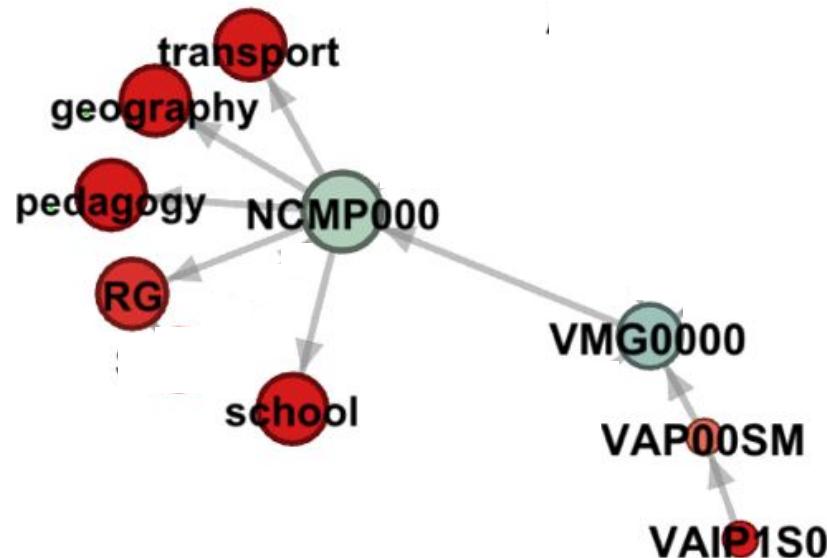


He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



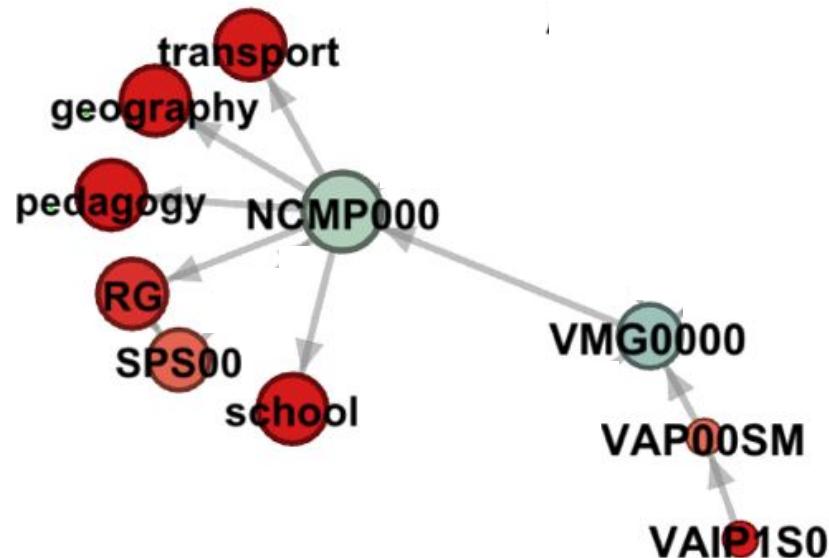
EmoGraph

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

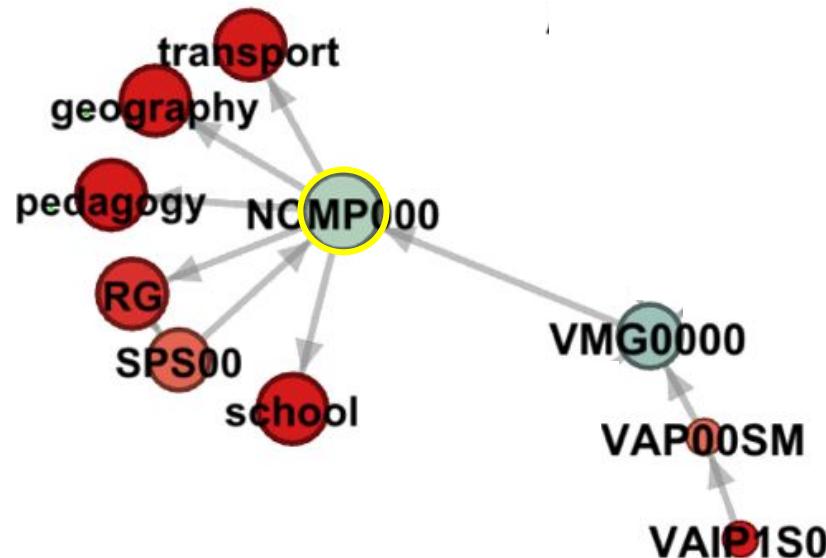


EmoGraph

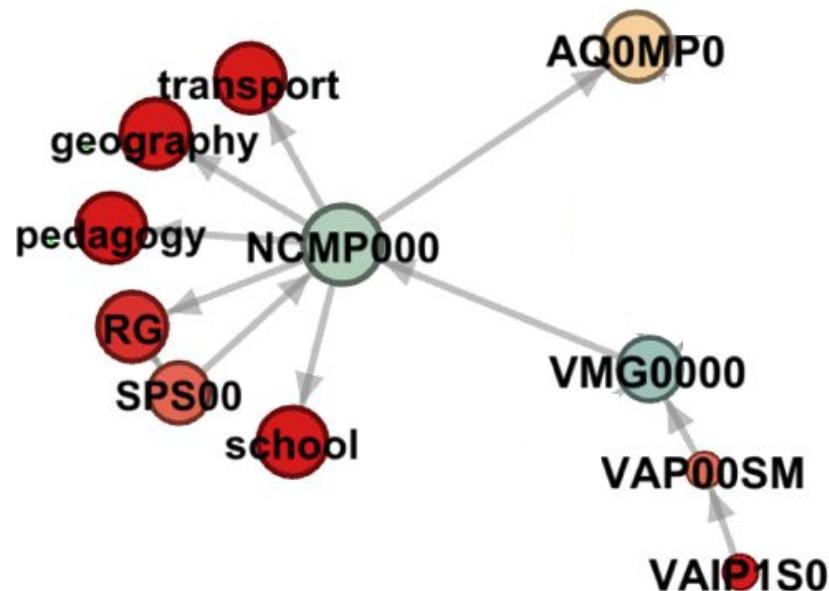
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



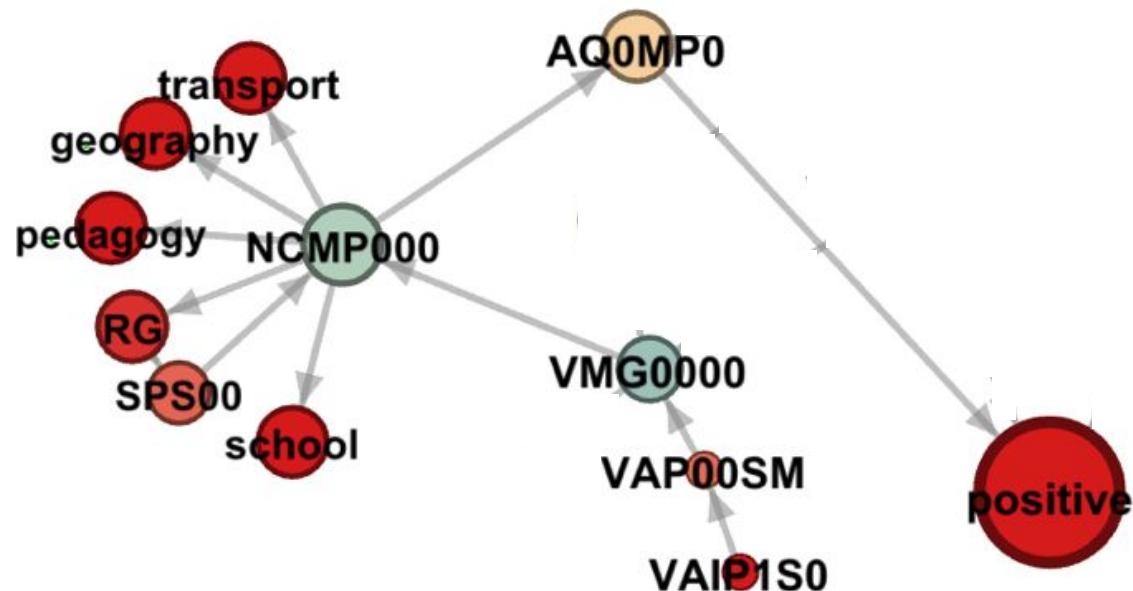
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



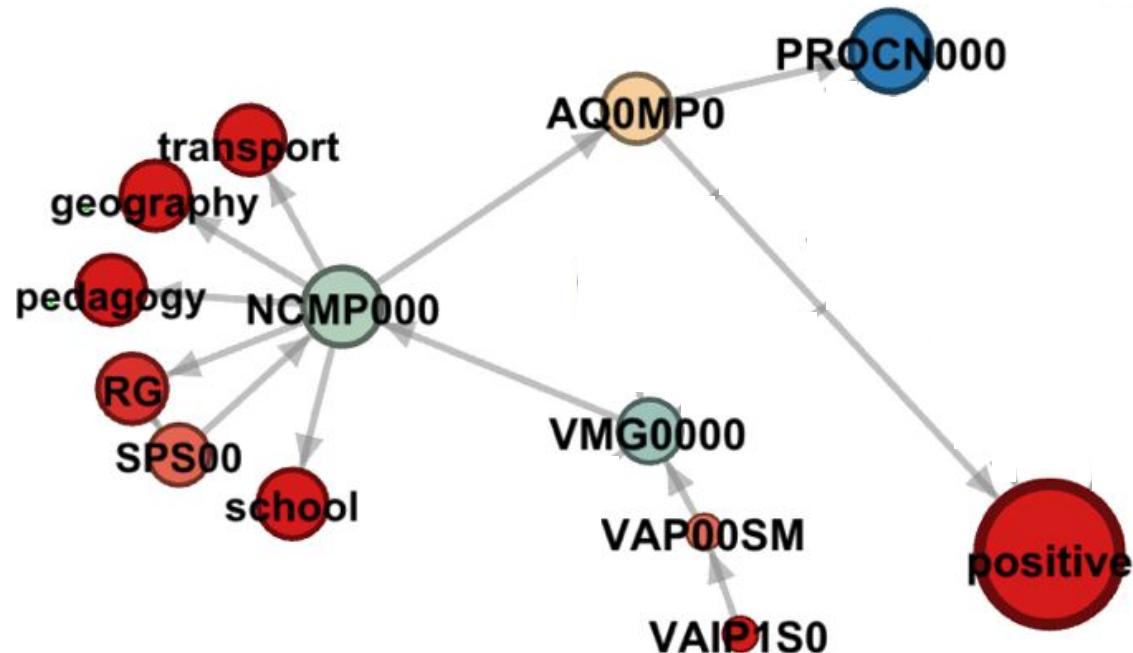
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



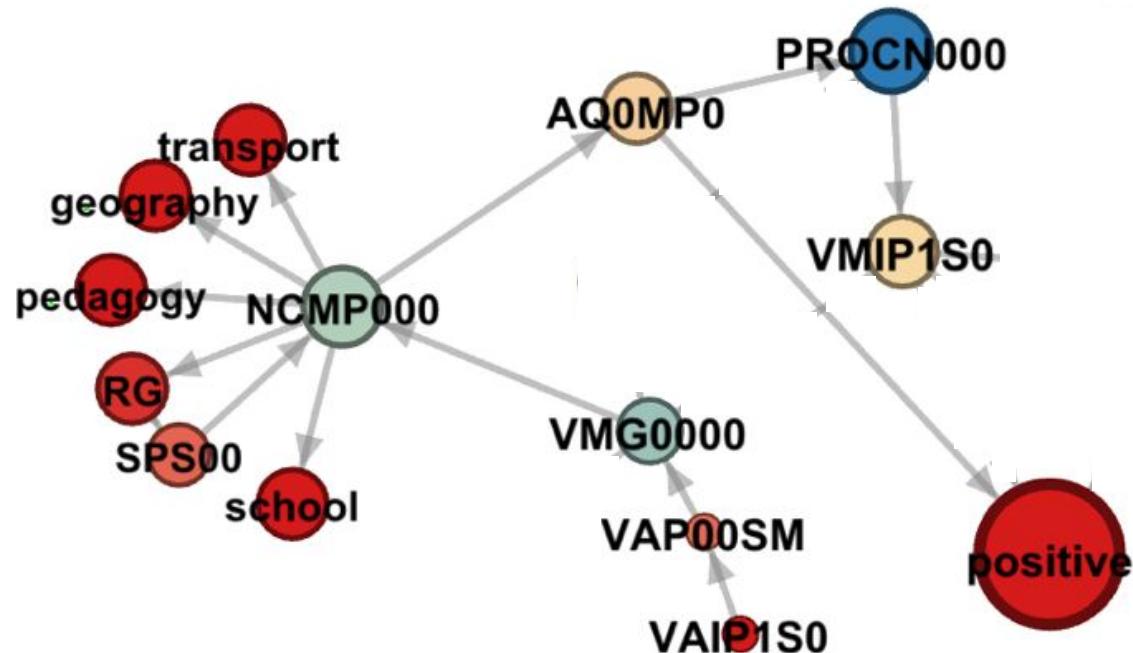
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



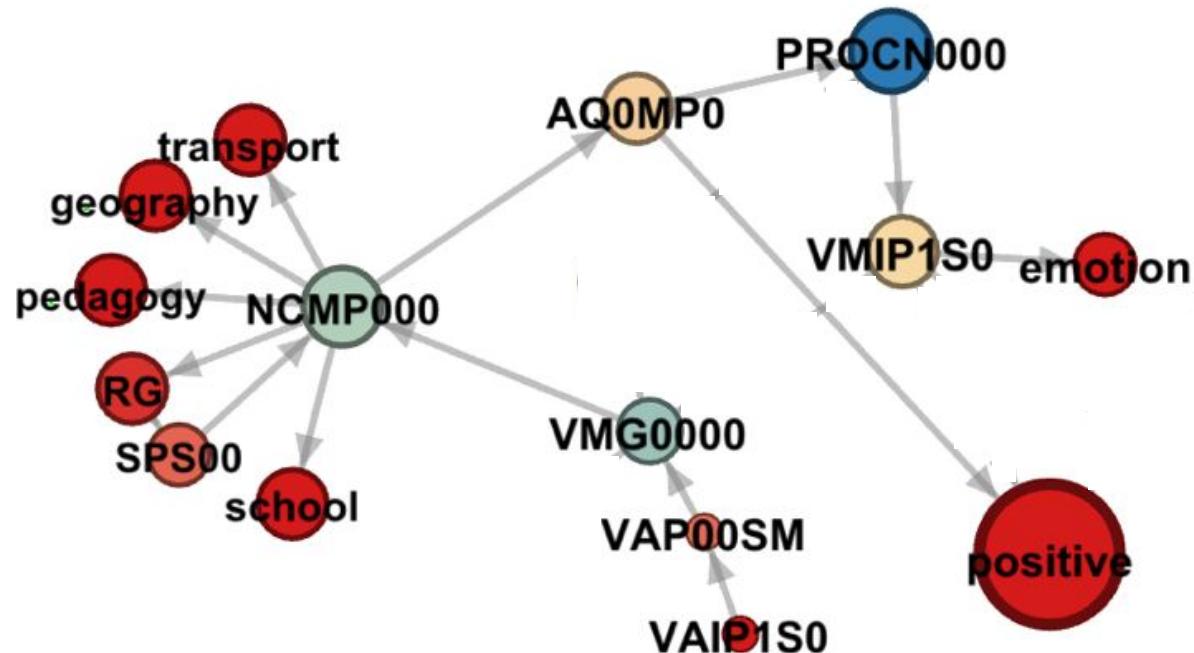
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



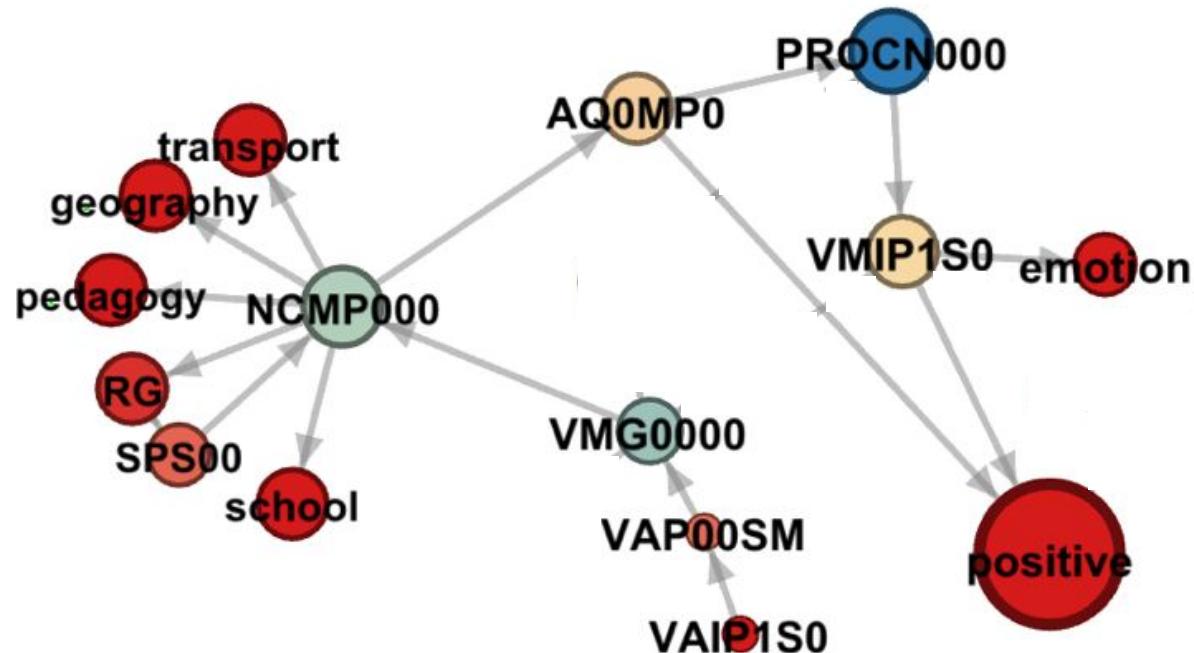
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



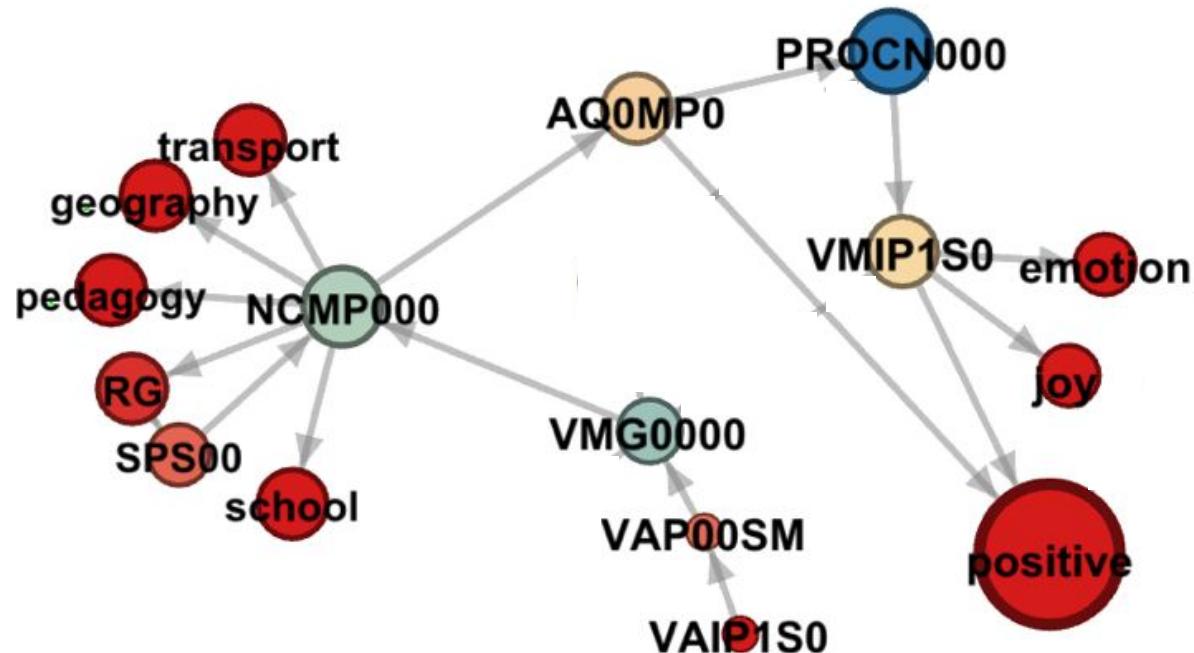
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



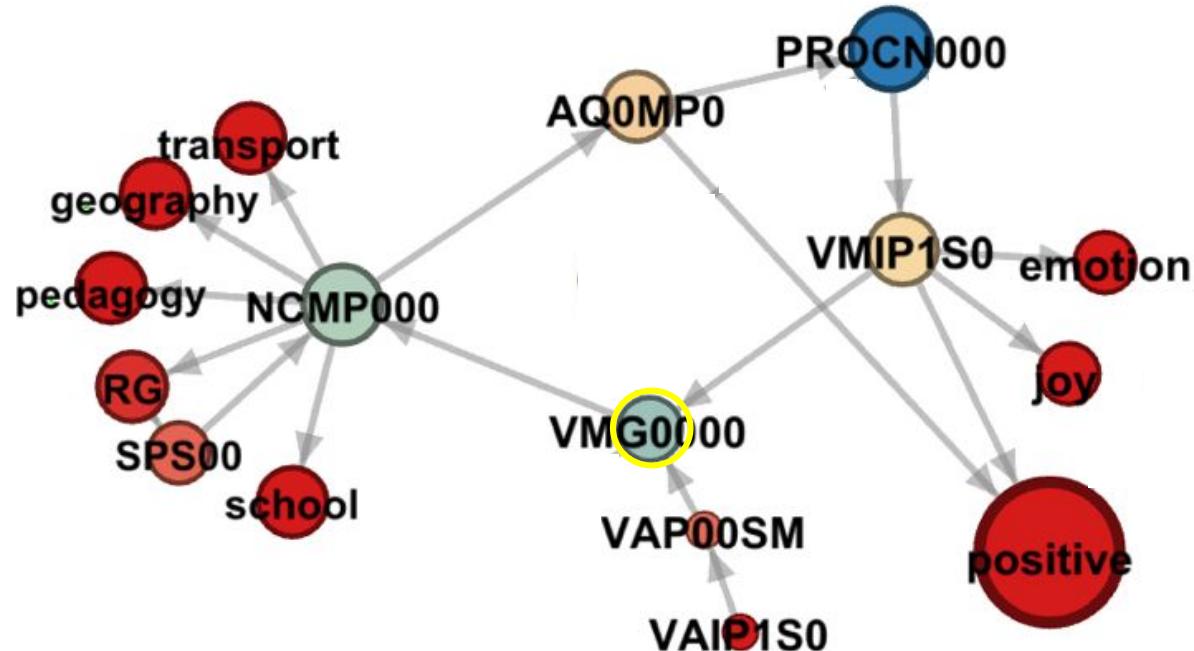
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



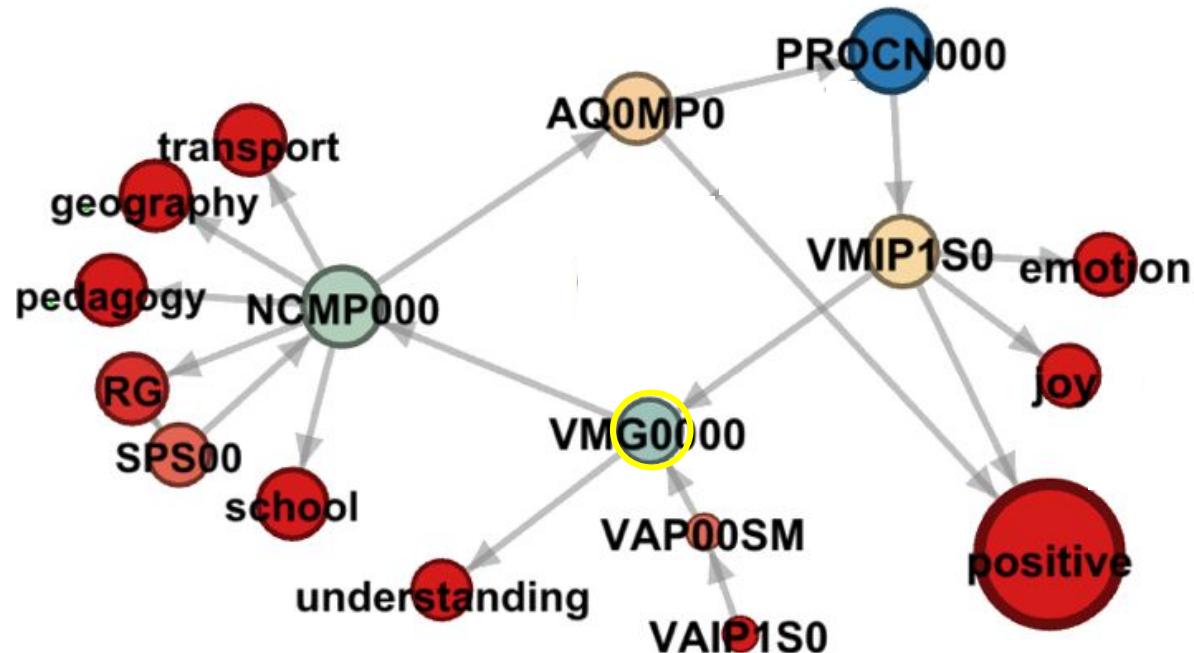
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



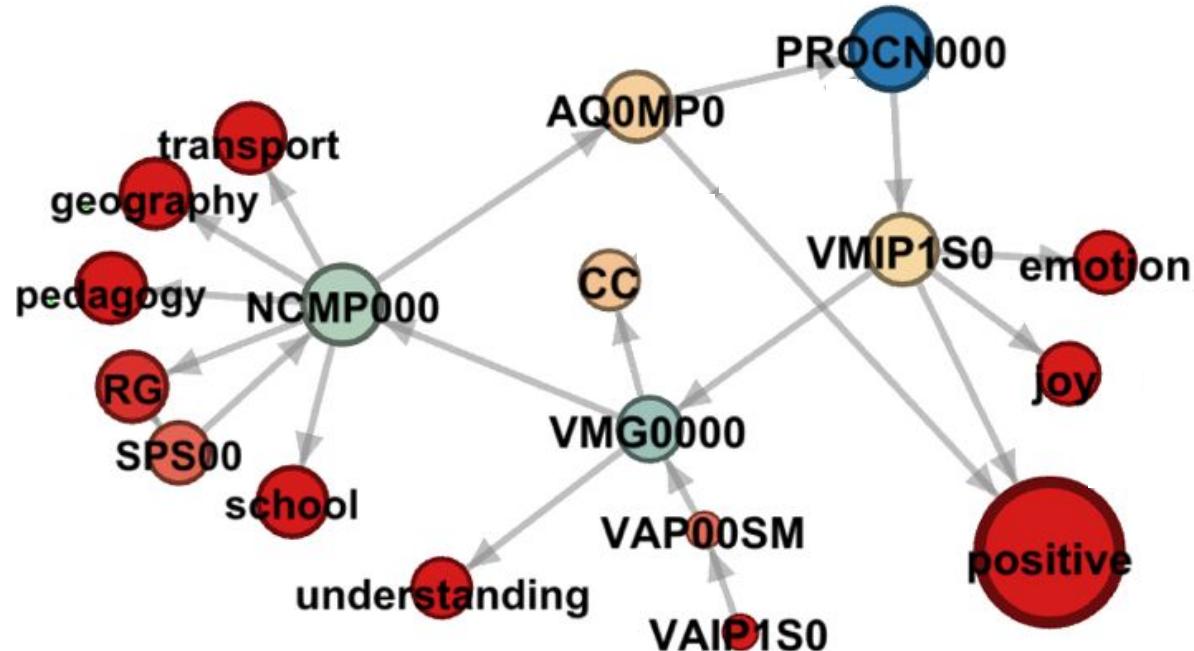
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



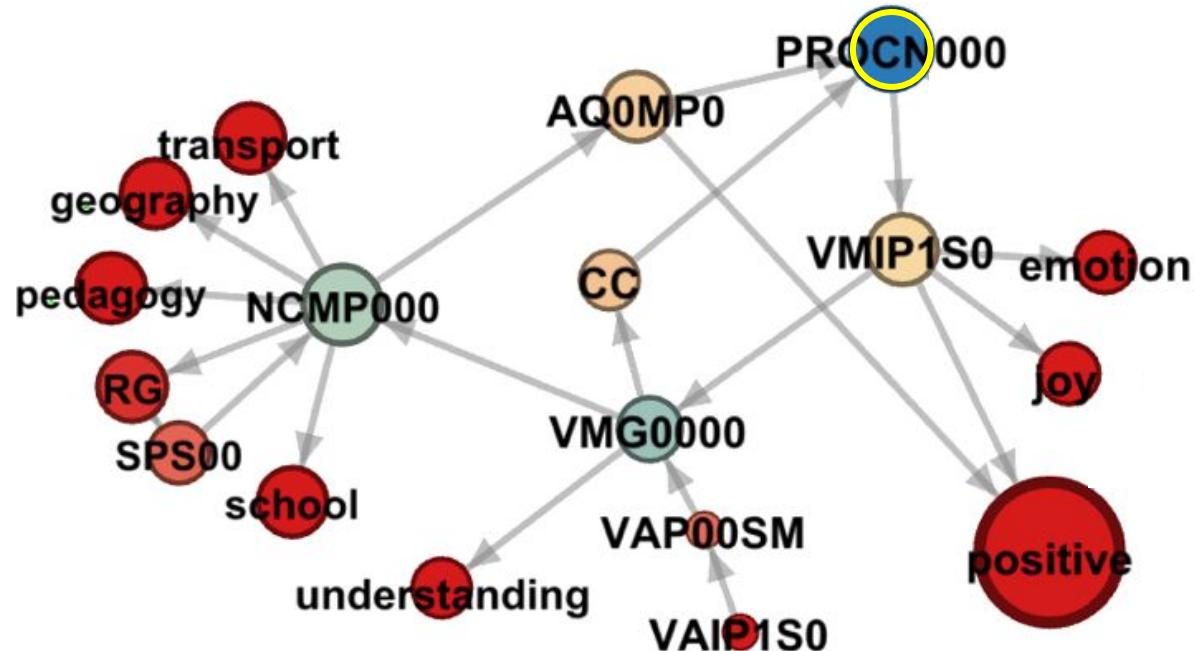
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



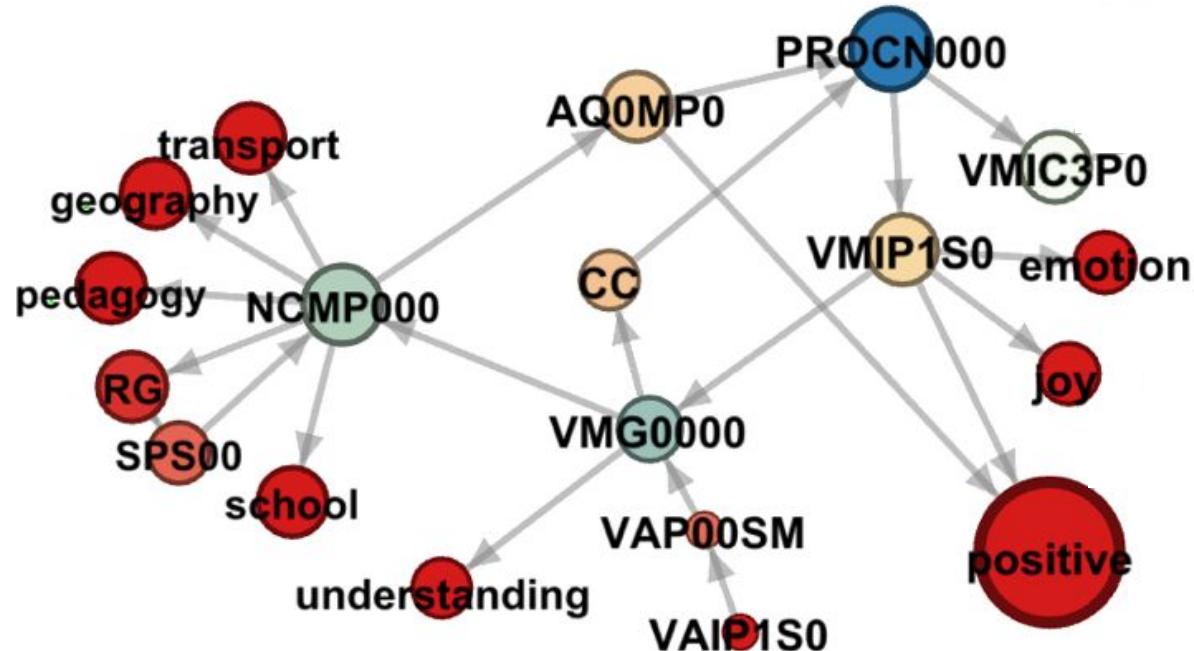
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



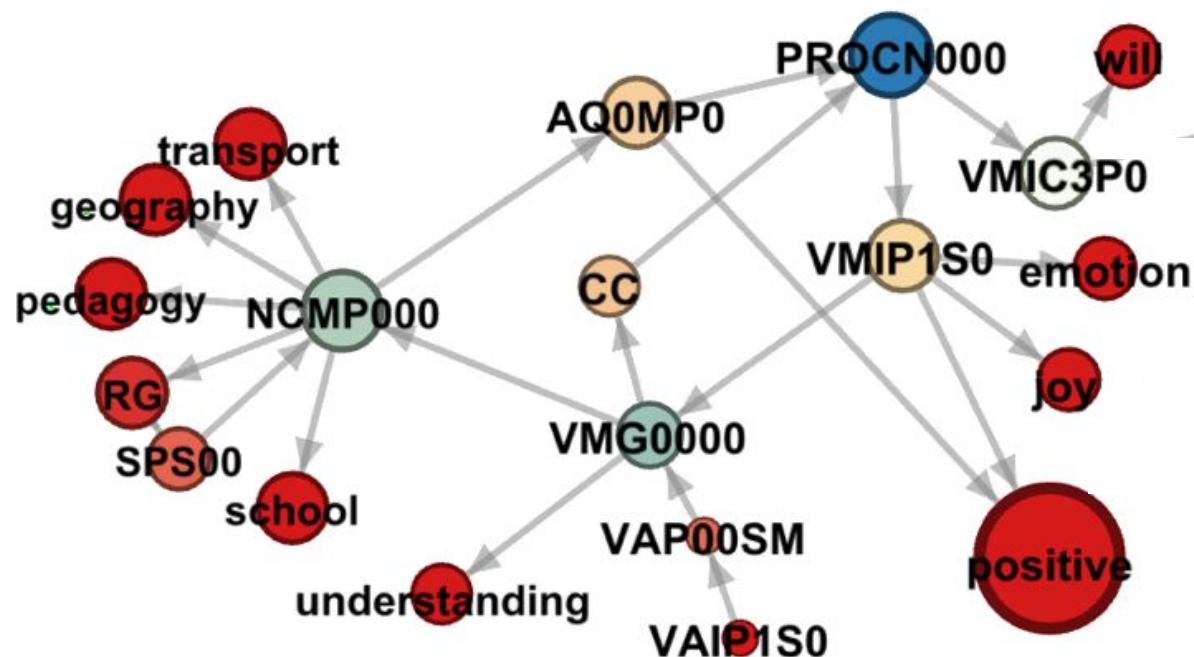
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



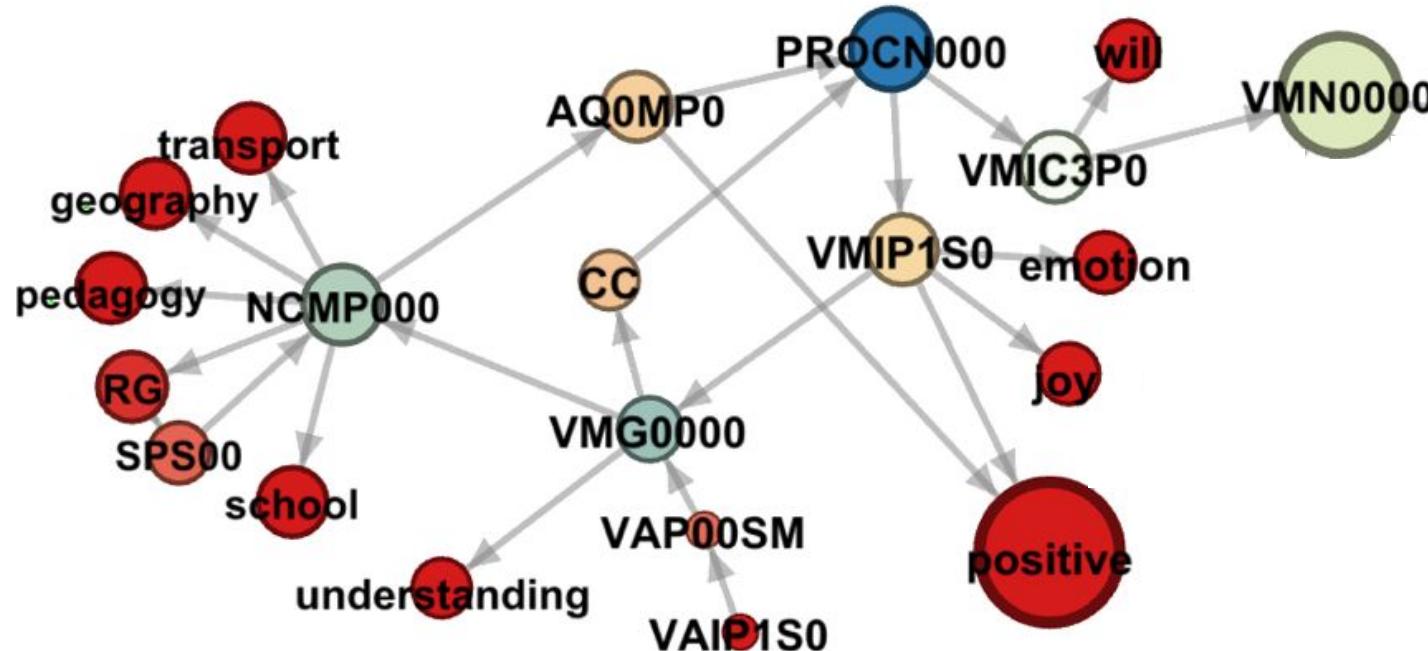
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



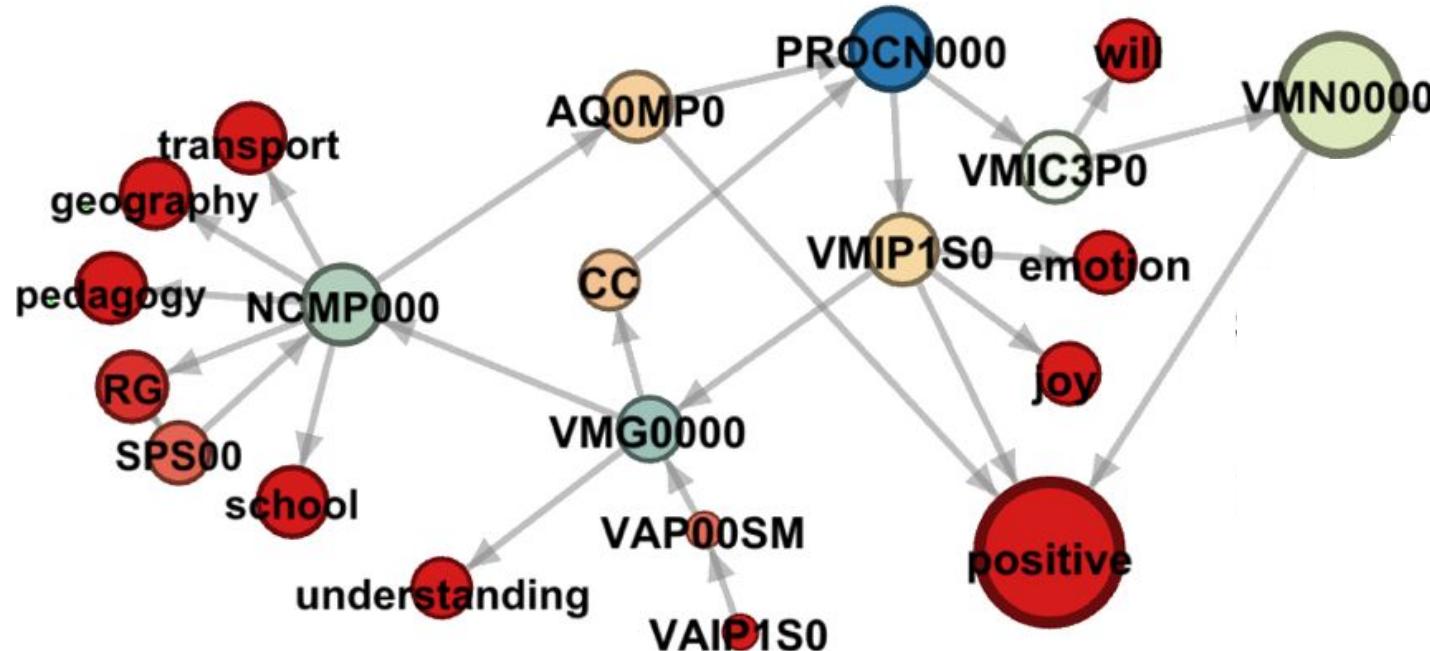
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



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

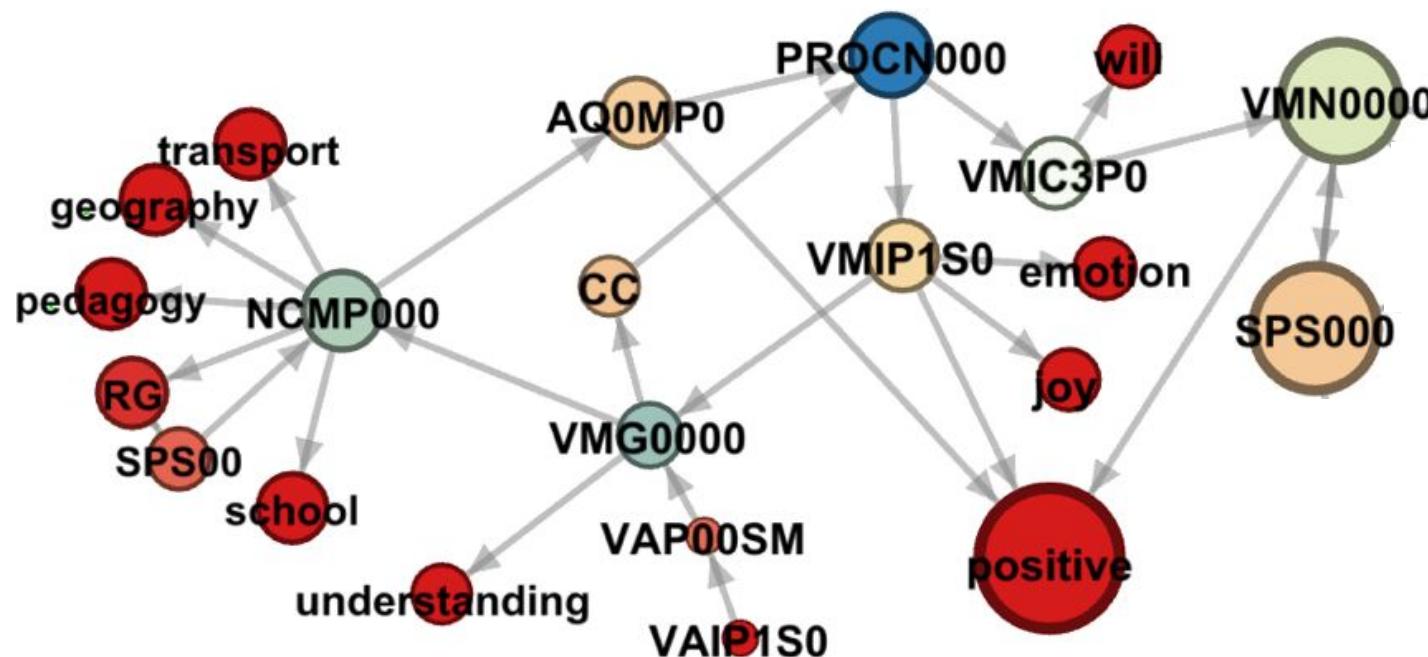


He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

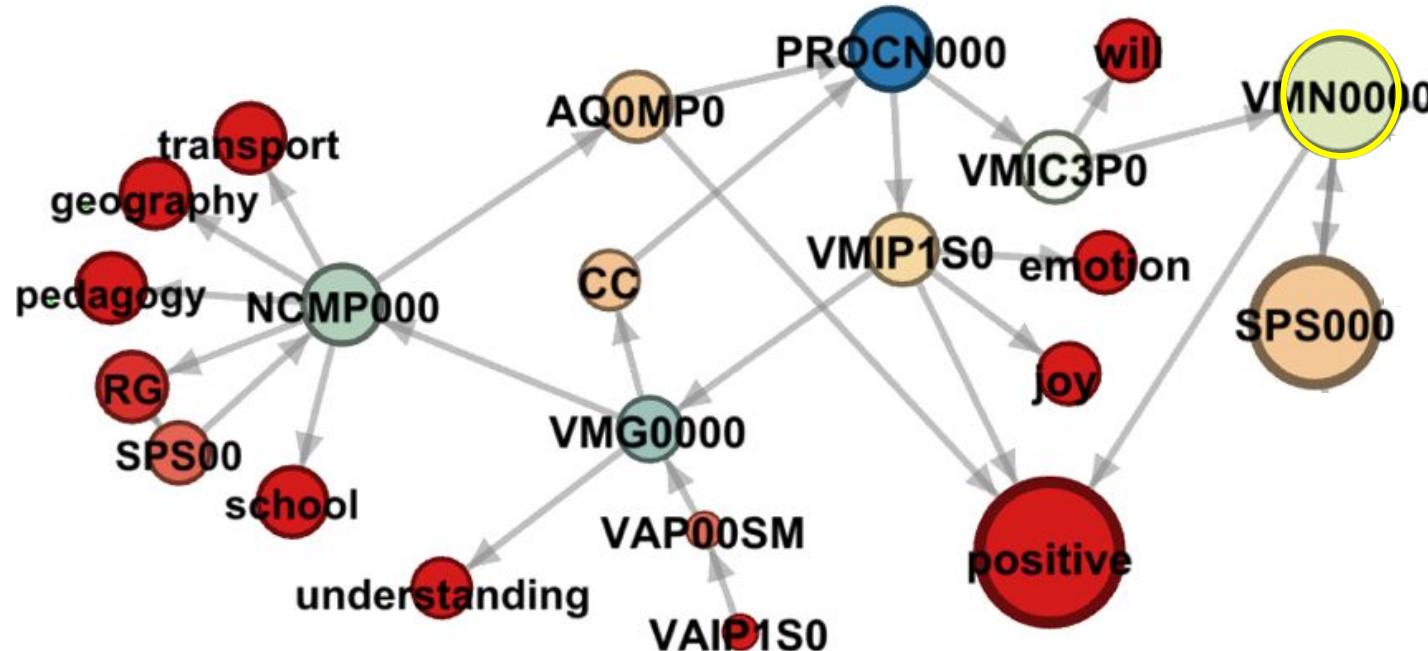


EmoGraph

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

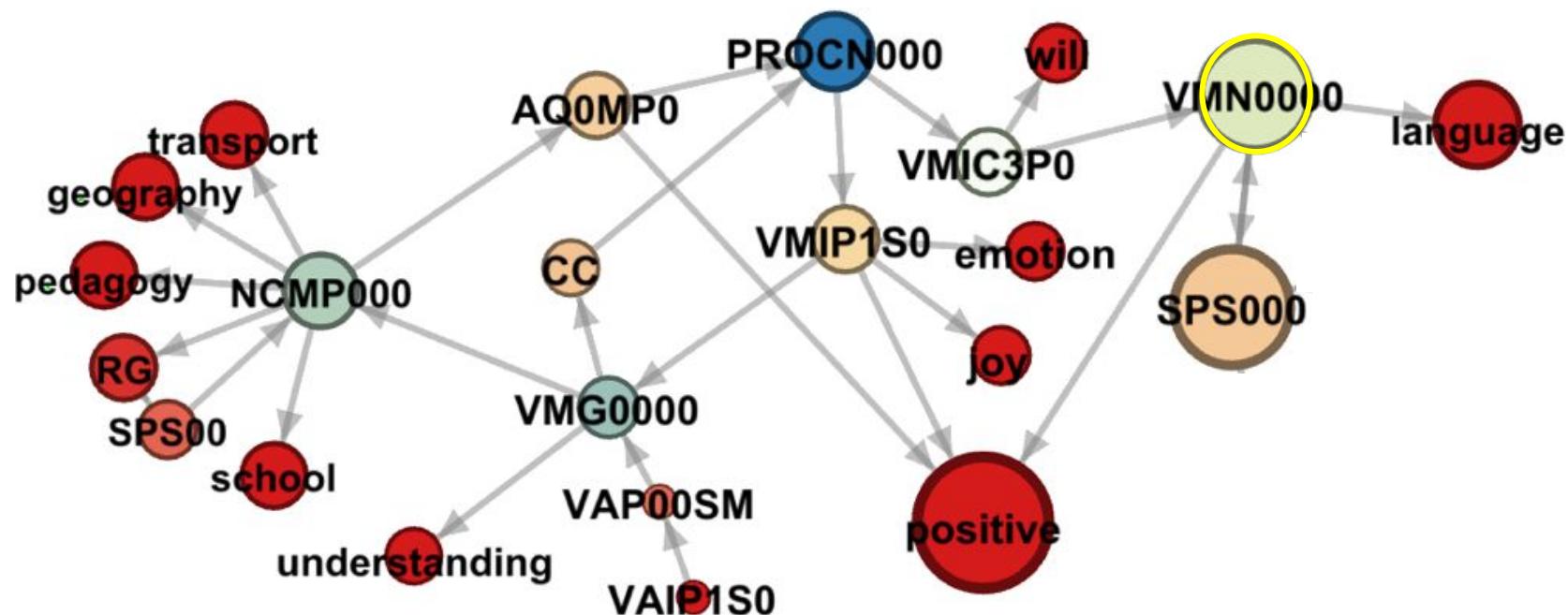


He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



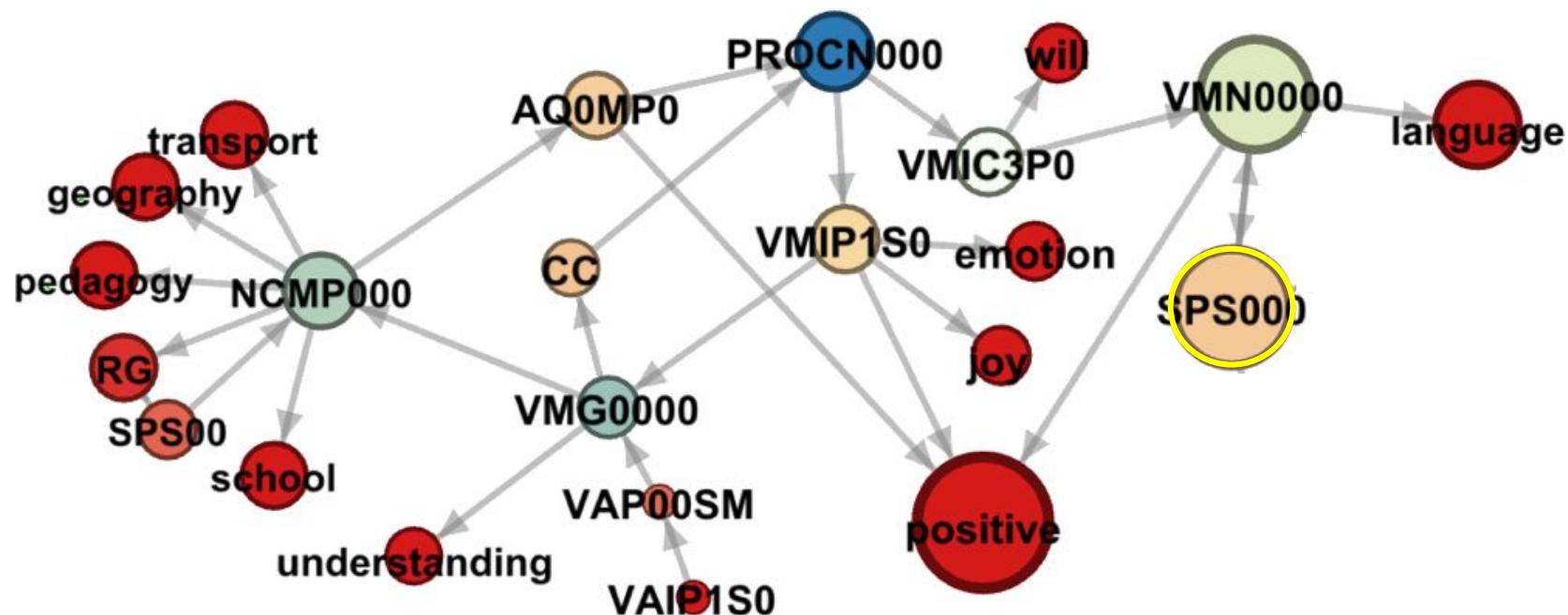
EmoGraph

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



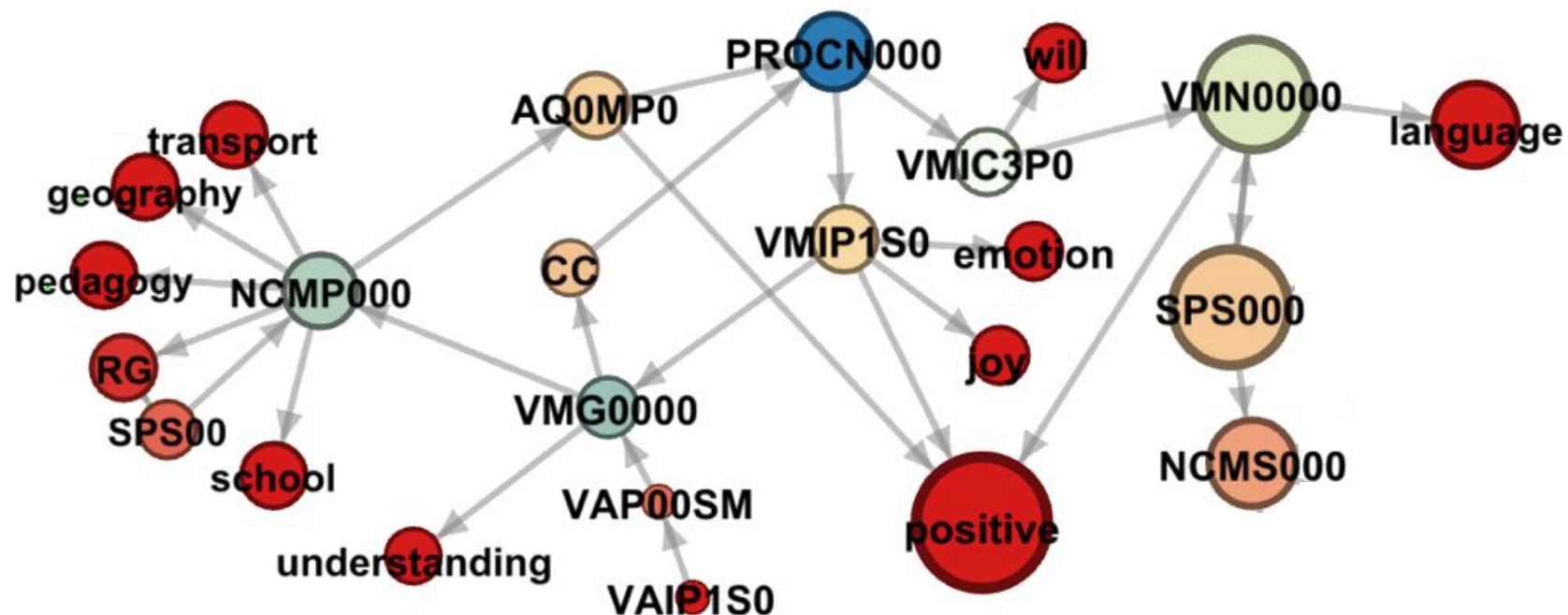
EmoGraph

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

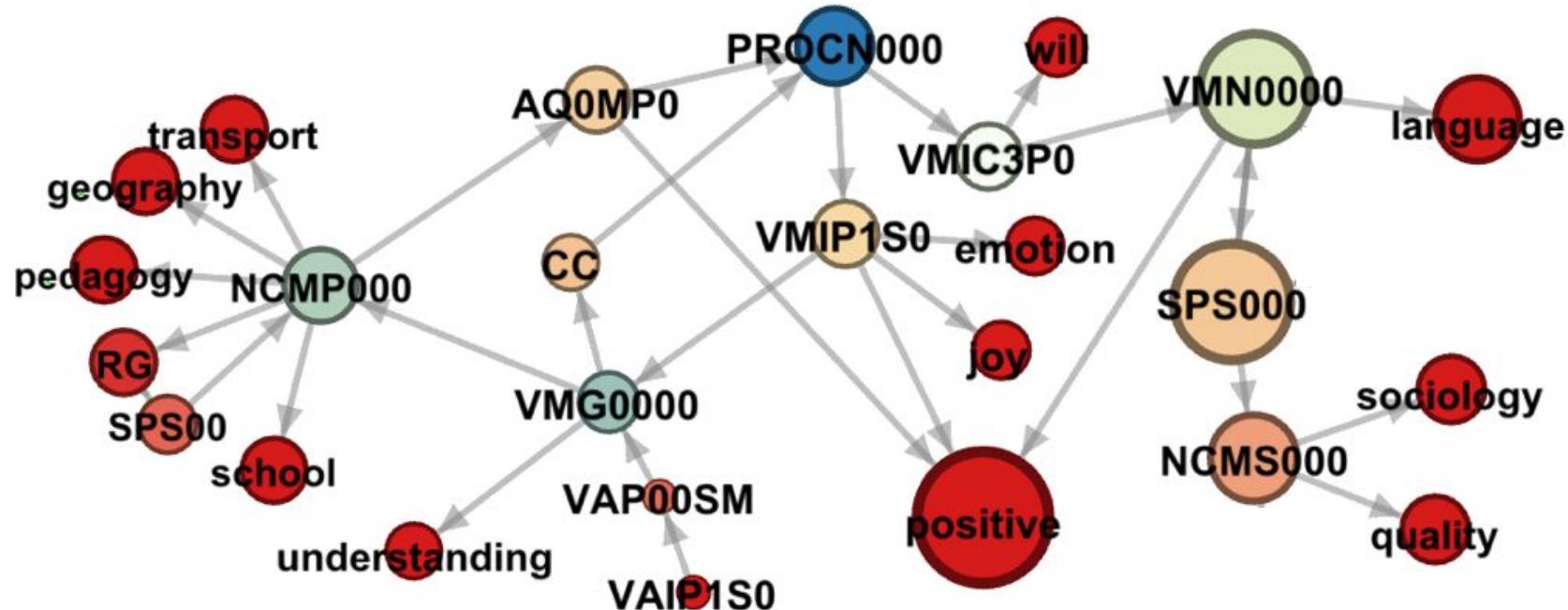


EmoGraph

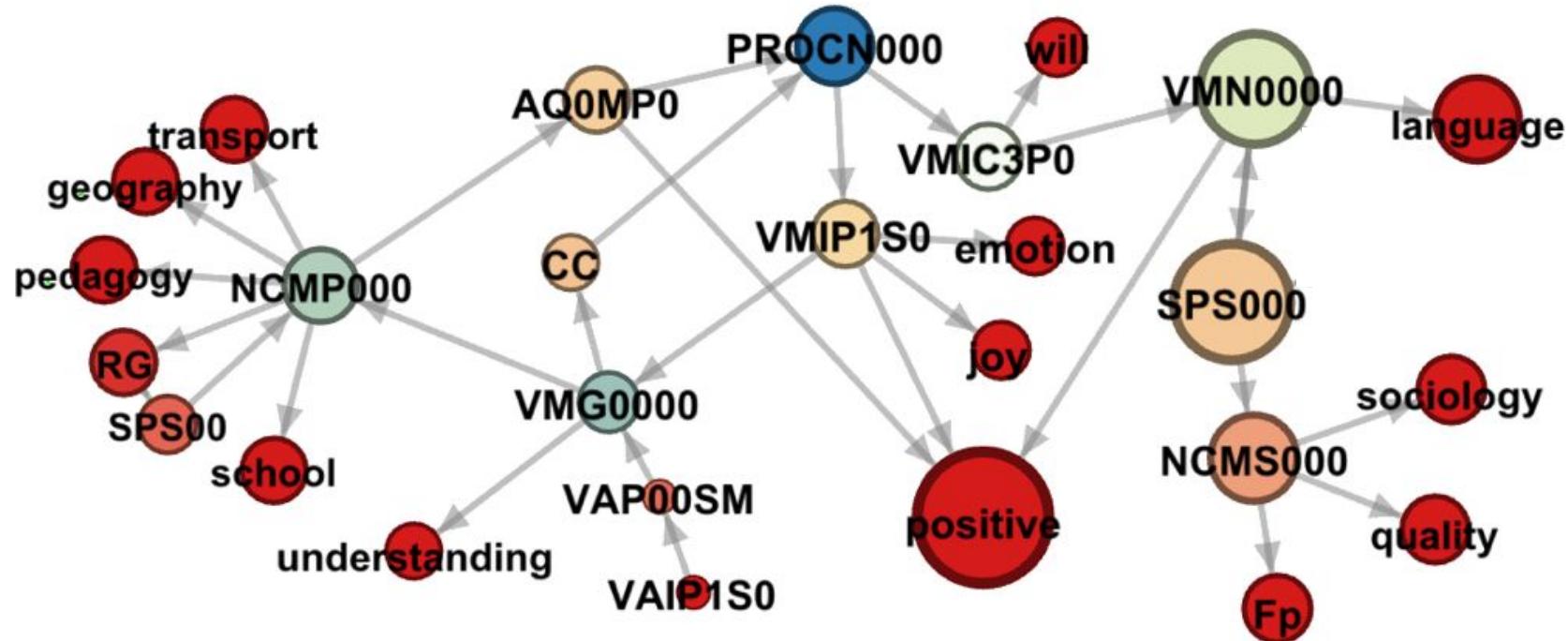
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.



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



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



Dado un grafo $G=\{N,E\}$ dónde,

- N es el conjunto de nodos
- E es el conjunto de ejes

Obtenemos dos conjuntos de características:

- características basadas en la estructura general del grafo
- características específicas de los nodos

Características del grafo

Nodos/ejes	Proporciona un indicador de conectividad del grafo. En nuestro caso, cómo de complejo es el discurso.	Máximo teórico: $\max(E) = N * (N - 1)$
Grado medio Grafo medio ponderado	Proporciona un indicador de cómo de interconectado está un grafo. En nuestro caso, cómo de interconectadas están las categorías gramaticales, tópicos y emociones entre sí.	Media de los grados de todos los nodos. Escalado entre [0,1]
Diámetro	Indica la mayor distancia entre cualquier par de nodos. En nuestro caso, cómo de lejos se encuentran las categorías gramaticales, tópicos y emociones entre sí.	$d = \max_{n \in N} \epsilon(N)$ donde E(N) es la eccentricidad
Densidad	Indica cómo de cerca está el grafo de ser completo. En nuestro caso, cómo de denso es el texto en el sentido de cómo cada categoría gramatical se usa en combinación con otras.	$D = \frac{2* E }{(N *(N -1))}$
Modularidad	Indica diferentes divisiones del grafo en grupos de nodos densamente conectados entre sí y poco conectados con nodos de otros módulos. En nuestro caso, cómo se modela el discurso en diferentes unidades estructurales y/o estilísticas.	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)
Coeficiente de agrupamiento	Indica la transitividad del grafo ($a \rightarrow b \ \&\& \ b \rightarrow c \mid P(a \rightarrow c)$). En nuestro caso, cómo se generan puentes entre categorías gramaticales, tópicos y emociones.	Watts-Strogatz: $cc1 = \frac{\sum_{i=1}^n C(i)}{n}$
Longitud media de camino	Indica cómo de lejos se pueden encontrar los nodos entre sí. En nuestro caso, cómo de lejos se encuentran algunas categorías gramaticales de otras, o cómo de lejos se encuentran las emociones que se expresan sobre los tópicos.	Brandes, U. A Faster Algorithm for Betweenness Centrality. In: Journal of Mathematical Sociology 25(2), pp. 163-177 (2001)

Características de los nodos

EigenVector (Vector Propio)	<p>Proporciona una medida de la importancia de cada nodo por el número de otros nodos importantes que lo enlazan.</p> <p>En nuestro caso, puede proporcionar las categorías gramaticales con un uso más central en el discurso. Por ejemplo, nombres, verbos o adjetivos.</p>	<p>Dado un grafo y su matriz de adyacencia $A = a_{n,t}$ donde $a_{n,t}$ es 1 si un nodo n está enlazado a un nodo t, y 0 en caso contrario:</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>donde λ es una constante que representa el mayor valor propio asociado con la medida de centralidad.</p>
Betweenness (Intermediación)	<p>Proporciona una medida de la importancia de cada nodo dependiendo del número de caminos más cortos de los cuales forma parte.</p> <p>En nuestro caso, si un nodo tiene un valor elevado de esta medida puede significar que es un conector común entre estructuras lingüísticas del discurso del autor. Por ejemplo, preposiciones, conjunciones o incluso verbos y nombres.</p>	<p>Es el ratio de todos los caminos más cortos entre cualquier par de nodos en el grafo que pasan por x:</p> $BC(x) = \sum_{i,j \in N - \{n\}} \frac{\sigma_{i,j}(n)}{\sigma_{i,j}}$ <p>Donde $\sigma_{i,j}$ es el número total de caminos más cortos del nodo i al nodo j, y $\sigma_{i,j}(n)$ es el número total de esos caminos que pasan por n.</p>

Representación vectorial EmoGraph

Estructura del grafo	8 características
ENRatio Degree WeightedDegree Diameter Density Modularity Clustering PathLength	Ratio nodos-arcos Grado medio del grafo Grado medio ponderado Diámetro del grafo Densidad del grafo Grado de modularidad Coeficiente de agrupamiento Longitud media del camino
Específicas de los nodos	944 características (472 nodos)
BTW-xx EIGEN-xx	Valor de intermediación (betweenness) de cada nodo (xx) Valor de vector propio (eigenvector) de cada nodo (xx)
Características Rangel-S	59 características

Corpus PAN-AP13

Introducción
Sexo y Edad
Variedad del Lenguaje
Conclusiones

Introducción
EmoGraph
Metodología
Resultados
Análisis y Discusión
Conclusiones

- Social Media en español. Gran tamaño, mucho ruido.

Edad*	Nº de Autores	
	Entrenamiento	Pruebas
10s	2.500	288
20s	42.600	4.608
30s	30.800	3.264
TOTAL	75.900	8.160

*Equilibrado por sexo

Aproximación

- **Características:**

- Rangel-S
- n-Gramas de POS (Rangel-nG)
- EmoGraph

- **Algoritmos de aprendizaje automático (Weka):**

Identificación de sexo	Máquinas de Vectores Soporte Núcleo Gausiano: g=0.20 c=1
Identificación de edad	Máquinas de Vectores Soporte Núcleo Gausiano: g=0.08 c=1

- **Medida de evaluación:**

- Accuracy
- t-Student ($H_0: p_1 = p_2$)

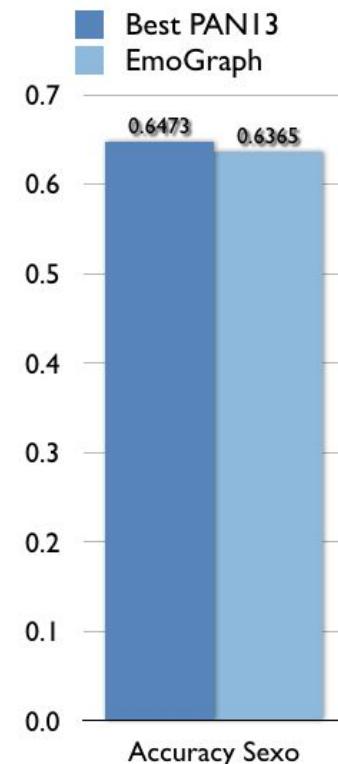
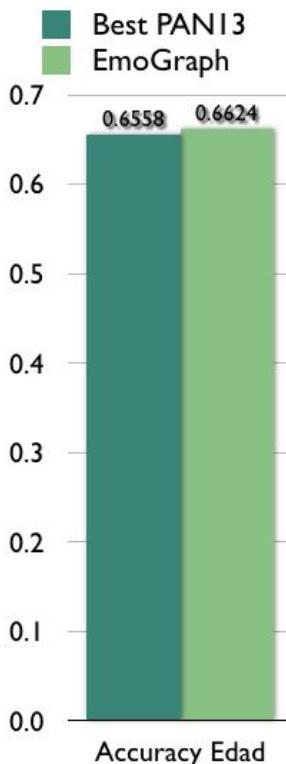
Identificación de Edad y Sexo

Introducción
Sexo y Edad
 Variedad del Lenguaje
 Conclusiones

Introducción
 EmoGraph
 Metodología
Resultados
 Análisis y Discusión
 Conclusiones

Pos.	Equipo	Accuracy
1	EmoGraph	66,24%
2	Pastor	65,58%
3	Santosh	64,30%
4	Rangel-S	63,50%
5	Haro	62,19%
6	Rangel-nG	61,62%
...		
21	Mechti	5,12%

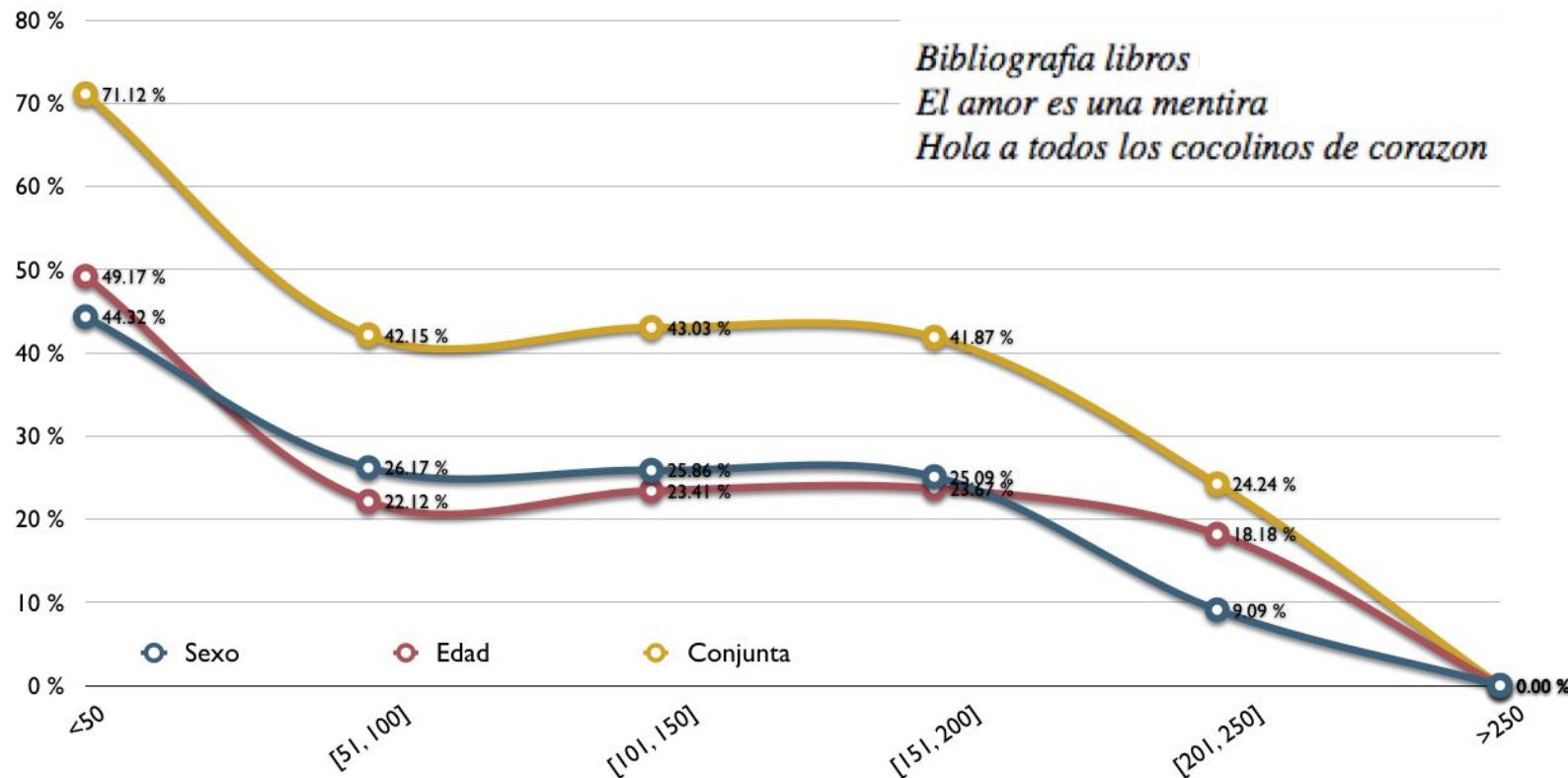
$$(z_{0.05} = 0.8894 < 1.960)$$



Pos.	Equipo	Accuracy
1	Santosh	64,73%
2	EmoGraph	63,65%
3	Pastor	62,99%
...		
7	Rangel-nG	60,16%
...		
9	Rangel-S	57,13%
...		
24	Gillam	47,84%

$$(z_{0.05} = 1.4389 < 1.960)$$

Análisis del error



Características más discriminantes

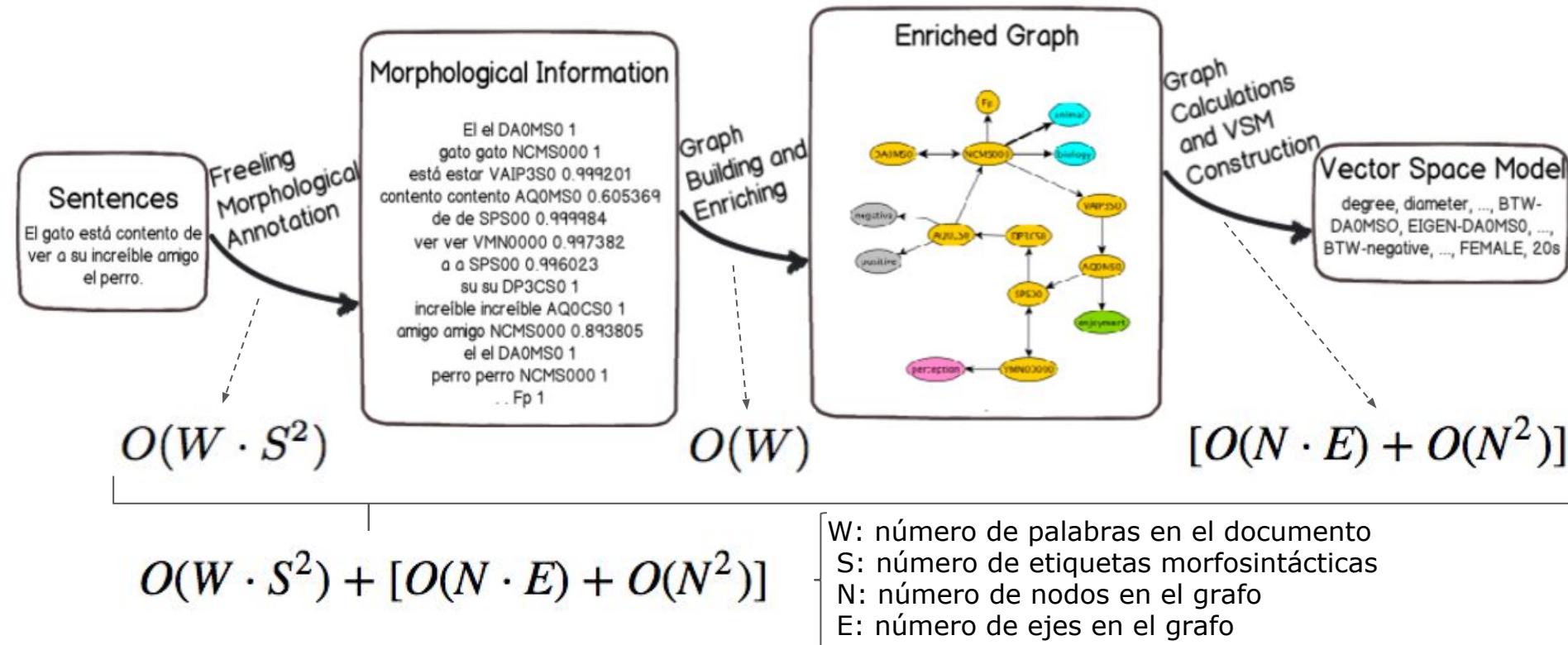
Ranking	Gender	Age	Ranking	Gender	Age
1	punctuation-seicolon	words-length	11	BTW-NC00000	EIGEN-SPS00
2	EIGEN-VMP00SM	Pron	12	BTW-Z	BTW-NC00000
3	EIGEN-Z	BTW-SPS00	13	EIGEN-DA0MS0	punctuation-exclamation
4	EIGEN-NCCP000	BTW-NCMS000	14	BTW-Fz	emoticon-happy
5	Pron	Intj	15	BTW-NCCP000	BTW-Fh
6	words-length	EIGEN-Fh	16	EIGEN-AQ0MS0	punctuation-colon
7	EIGEN-NC00000	BTW-PP1CS000	17	SEL-disgust	punctuation
8	EIGEN-administration	EIGEN-Fpt	18	EIGEN-DP3CP0	BTW-Fpt
9	Intj	EIGEN-NC00000	19	EIGEN-DP3CS0	EIGEN-DA0FS0
10	SEL-sadness	EIGEN-NCMS000	20	SEL-anger	Verb

- Características eigen en sexo vs. betweenness en edad.
- Verbos, nombres y adjetivos en sexo vs. preposiciones y signos de puntuación en edad.
- Alta presencia de características de emoción en identificación de sexo.

Análisis de costes

Introducción
Sexo y Edad
Variedad del Lenguaje
Conclusiones

Introducción
EmoGraph
Metodología
Resultados
Análisis y Discusión
Conclusiones



Robustez frente a medios e idiomas

- Corpus PAN-AP14 - Múltiples medios. Inglés y español.

Edad*	Social Media		Blogs		Twitter		Revisiones
	Inglés	Español	Inglés	Español	Inglés	Español	Inglés
18-24	680	150	10	4	12	4	74
25-34	900	180	24	12	56	26	200
35-49	980	138	32	26	58	46	200
50-64	790	70	10	10	26	12	200
65+	26	28	2	2	2	2	147
TOTAL	3.376	566	78	54	154	90	821

(Rangel & Rosso, CLEF 2015)

*Equilibrado por sexo. Se muestra sólo test

Robustez frente a medios e idiomas

- **Características:**
 - 6-gramas caracteres (1.000 más frecuentes) + EmoGraph
- **Algoritmos de aprendizaje automático (Weka):**

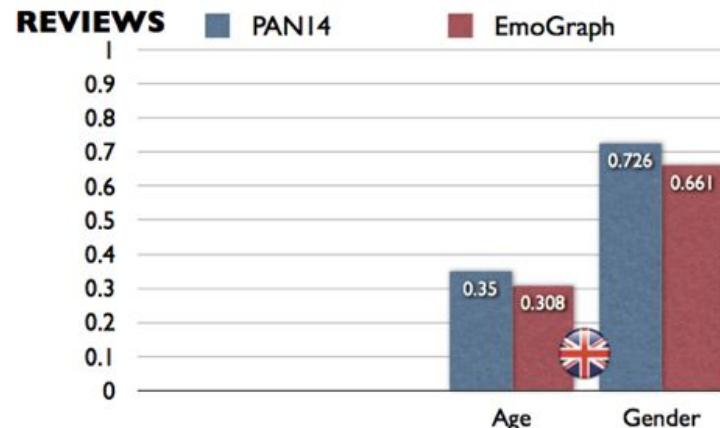
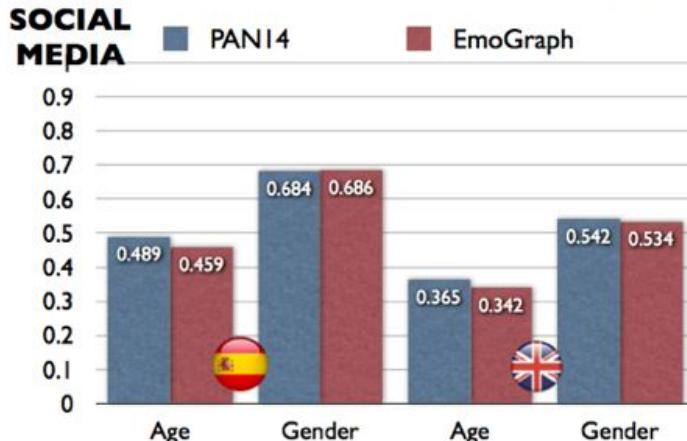
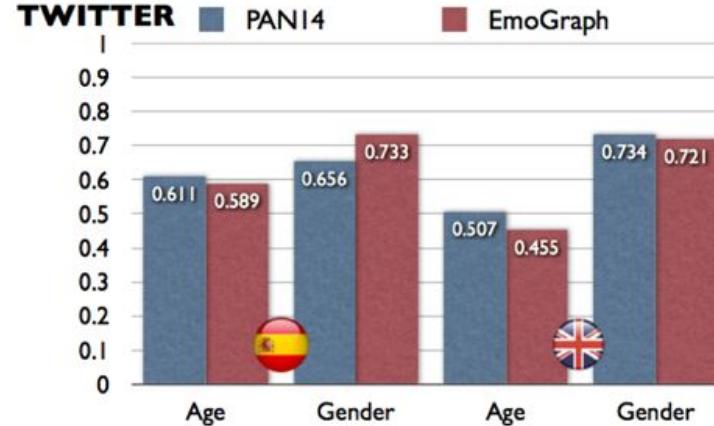
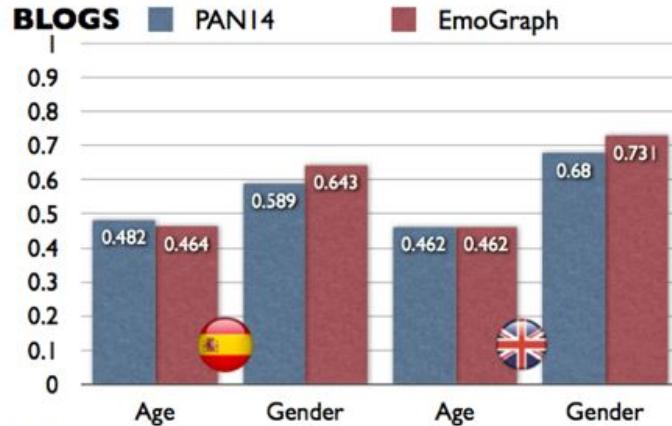
Identificación de sexo Twitter inglés	Regresión Logística
Identificación de sexo y edad revisiones y social media en inglés	Máquinas de Vectores Soporte
Identificación de edad Twitter español	Máquinas de Vectores Soporte
El resto	Adaboost (Decision Stump)

- **Medida de evaluación:**
 - Accuracy

Robustez frente a medios e idiomas

Introducción
Sexo y Edad
Variedad del Lenguaje
Conclusiones

Introducción
EmoGraph
Metodología
Resultados
Análisis y Discusión
Conclusiones



Contribución de EmoGraph (EmIroGeFB)

Introducción
Sexo y Edad
Variedad del Lenguaje
Conclusiones

Introducción
EmoGraph
Metodología
Resultados
Análisis y Discusión
Conclusiones

Características	Accuracy
F + P	52,92%
C	55,42%
F + P + C	56,25%
F + P + C + E + SEL	59,09%
Simple Graph	50,83%
Complete Graph	51,92%
Semantic Graph	55,01%
EmoGraph	65,96%

LEYENDA

- F: Frecuencias
- P: Signos de puntuación
- C: Categorías gramaticales
- E: Emoticonos
- SEL: Lexicón emociones
- Simple Graph: Grafo categorías gramaticales
- Complete Graph: Simple + info. morfosintáctica
- Semantic Graph: Complete + info. semántica
- EmoGraph: Semantic + emociones

$$z_{0.05} = 3.4764 >> 1.960$$

Temas por sexo (PAN-AP13)

MUJERES



Introducción
Sexo y Edad
Variedad del Lenguaje
Conclusiones

Introducción
EmoGraph
Metodología
Resultados
Análisis y Discusión
Conclusiones

HOMBRES



- Sin diferencias significativas entre sexos.
- Se habla de la vida, el amor y la esperanza.

Evolución de temas por sexo y edad (PAN-AP13)

Introducción *Sexo y Edad*

Variedad del Lenguaje Conclusiones

Introducción
EmoGraph
Metodología
Resultados
Análisis y Discusión
Conclusiones

MUJERES



10s

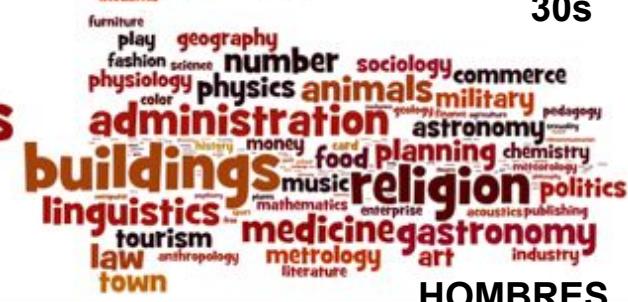


A word cloud visualization where the size and color of words represent their frequency or importance. The most prominent words include 'medicine' (large, red), 'buildings' (large, green), 'gastronomy' (large, blue), 'religion' (large, orange), 'law' (medium, yellow), 'linguistics' (medium, pink), 'music' (medium, purple), 'tourism' (medium, red), 'commerce' (medium, blue), 'gastronomy' (medium, blue), 'metrology' (medium, yellow), 'mathematics' (medium, blue), 'sexuality' (medium, blue), 'biology' (medium, green), 'geology' (medium, green), 'chemistry' (medium, green), 'physics' (medium, blue), 'astronomy' (medium, red), 'planetary science' (medium, blue), 'fashion' (medium, red), 'industry' (medium, blue), 'planning' (medium, red), 'psychiatry' (medium, blue), 'cardiology' (medium, blue), 'literature' (medium, blue), 'number' (medium, blue), 'acoustics' (medium, blue), 'geography' (medium, blue), 'crown' (medium, blue), 'animals' (medium, blue), 'physiology' (medium, green), and 'color' (medium, green).

20s



30s



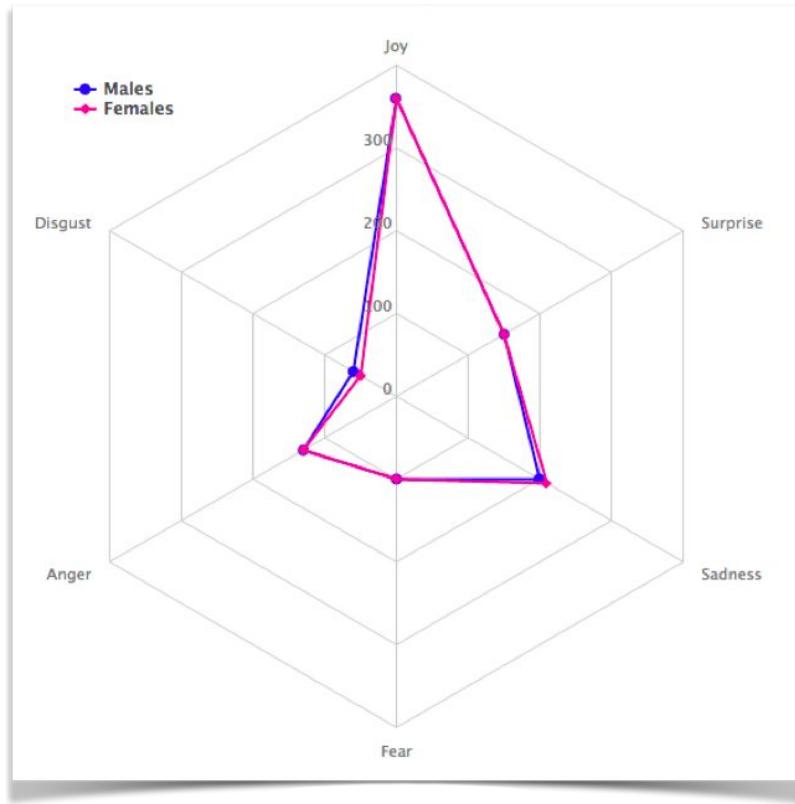
HOMBRES

- Los más jóvenes hablan de estudios: (hombres) physics, law... (mujeres) chemistry, linguistics...
 - Según crecemos nos interesamos por la vivienda (buildings), animales, gastronomía, medicina, y religión.
 - Las adolescentes hablan más de sexo que los adolescentes que hablan de compras (shopping).

Emociones por sexo (PAN-AP13)

Introducción
Sexo y Edad
Variedad del Lenguaje
Conclusiones

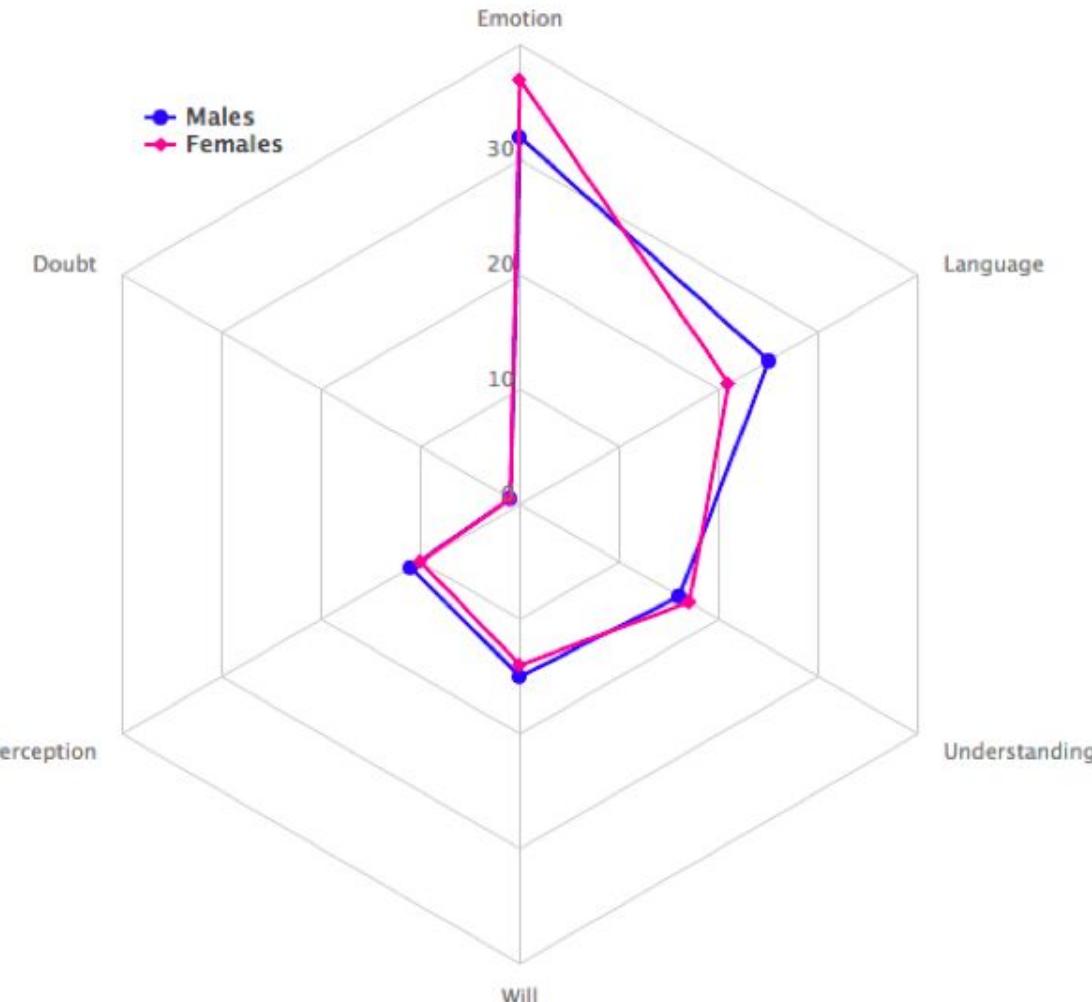
Introducción
EmoGraph
Metodología
Resultados
Análisis y Discusión
Conclusiones



- Sin diferencias significativas entre sexos.
- Las mujeres parecen expresar ligeramente más disgusto que los hombres.
- Los hombres por su parte lo hacen con tristeza.

Tipos de verbos por sexo (PAP13)

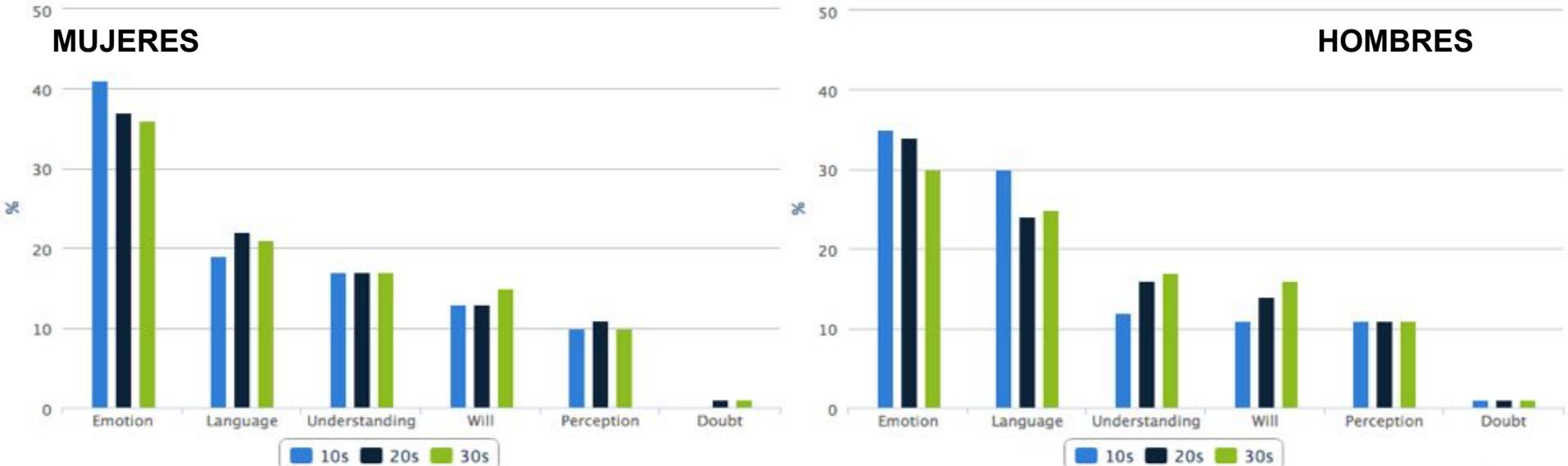
- Las mujeres usan más verbos de emoción (sentir, querer, amar...).
- Los hombres usan más verbos relativos al lenguaje (decir, contar, hablar...).



Evolución t. de verbos por sexo (PAN-AP13)

Introducción
Sexo y Edad
Variedad del Lenguaje
Conclusiones

Introducción
EmoGraph
Metodología
Resultados
Análisis y Discusión
Conclusiones



- El uso de verbos de emoción decrece con los años.
- Las mujeres usan más verbos de emoción a cualquier edad, mientras que los hombres lo hacen con verbos relativos al lenguaje.
- A diferencia de las mujeres, los hombres incrementan el uso de verbos de entendimiento con los años. En su edad adulta, los usan al mismo nivel que las mujeres en su adolescencia.

- EmoGraph (grafos + emociones) permite corroborar nuestra hipótesis:

discurso + emociones --> edad y sexo
 - Independiente del **medio social** y del **idioma**.
- En línea con los estudios referentes:
 - El **sexo** difiere en el modo de organizar el discurso en torno a determinadas **categorías gramaticales, temas y emociones**.
 - La **edad** difiere en el modo de articular el discurso, según el modo de **conectar** sus diferentes **elementos**.
- EmoGraph requiere de un **mínimo de palabras** para funcionar de manera óptima.
- Su complejidad computacional permite aplicarlo a entornos **Big Data** como son los **medios sociales**.

¿Nos diferenciamos los hablantes de una lengua, en sus distintos dialectos o variedades, al momento de expresar nuestras emociones?, ¿o las variaciones se producen a otro nivel, como por ejemplo en las palabras usadas para hacerlo?

English	I was goofing around with my dog and I lost my mobile .
ES-Argentina	Estaba haciendo boludeces con mi perro y extravié el celular .
ES-Mexico	Estaba haciendo el pendejo con mi perro y extravié el celular .
ES-Spain	Estaba haciendo el tonto con mi perro y perdí el móvil .

emociones (EmoGraph) vs. contenido

(Rangel et al., CICLING 2016)

Motivación: En los **medios sociales no existen fronteras** geográficas. ¿Cómo podemos **segmentar** por regiones? (no sólo marketing, también lingüística forense/seguridad).

- ¿Cómo se distribuye geográficamente la opinión pública?
- ¿Qué influencias culturales tiene el autor de una nota de amenaza?

El problema puede consistir discriminar entre **variedades de una misma lengua**.

Se considera una tarea de **author profiling**: influencia de la **idiosincrasia cultural** en el autor (e.g. diferentes expresiones, vocabulario...).

Restricción: **Big Data -> Low Dimensionality Representation (LDR)**

Procedimiento formal

Paso 1. Matriz de pesos tf-idf de los términos de los documentos:

$$\Delta = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1m} & \delta(d_1) \\ w_{21} & w_{22} & \dots & w_{2m} & \delta(d_2) \\ \dots & \dots & \dots & \dots & \dots \\ w_{n1} & w_{n2} & \dots & w_{nm} & \delta(d_n) \end{bmatrix}$$

- Cada columna es un término t del vocabulario
- Cada fila representa a un documento d
- w_{ij} es el peso tf-idf del término j en el documento i
- $\delta(d_i)$ representa la clase c asignada al documento i

Paso 2. Peso de los términos dependiente de la clase:

$$W(t, c) = \frac{\sum_{d \in D / c = \delta(d)} w_{dt}}{\sum_{d \in D} w_{dt}}, \forall d \in D, c \in C$$

Paso 3. Representación de los documentos dependiente de la clase:

$$d = \{F(c_1), F(c_2), \dots, F(c_n)\} \sim \forall c \in C, \quad F(c_i) = \{avg, std, min, max, prob, prop\}$$

Cálculo de las medidas

avg	El peso medio de un documento se calcula como la suma de todos los pesos $W(t,c)$ de sus términos, dividido por el número total de términos del vocabulario en el documento.
std	La desviación estándar de los pesos de un documento se calcula como la raíz cuadrada de la suma de todos los términos $W(t,c)$ menos la media.
min	El peso mínimo de un documento es el menor valor de los pesos $W(t,c)$ del documento.
max	El peso máximo de un documento es el mayor valor de los pesos $W(t,c)$ del documento.
prob	El peso global de un documento es la suma de pesos $W(t,c)$ de los términos del documento dividido por el número total de términos del documento.
prop	Proporción entre el número de términos del vocabulario presentes en el documento y el número total de términos en el documento.

Variedad	# Blogs (autores)	
	Entrenamiento	Pruebas
AR-Argentina	450	200
CL - Chile	450	200
ES - España	450	200
MX - México	450	200
PE - Perú	450	200
TOTAL	2.250	1.000

- Autores censados por expertos de Autoritas nativos y residentes en los diferentes países.
- No se comparten autores entre entrenamiento y pruebas, para reducir posible sobreajuste.

<https://github.com/autoritas/RD-Lab/tree/master/data/HispaBlogs>

(Franco et al., DSL 2015; Rangel et al., CICLING 2016)

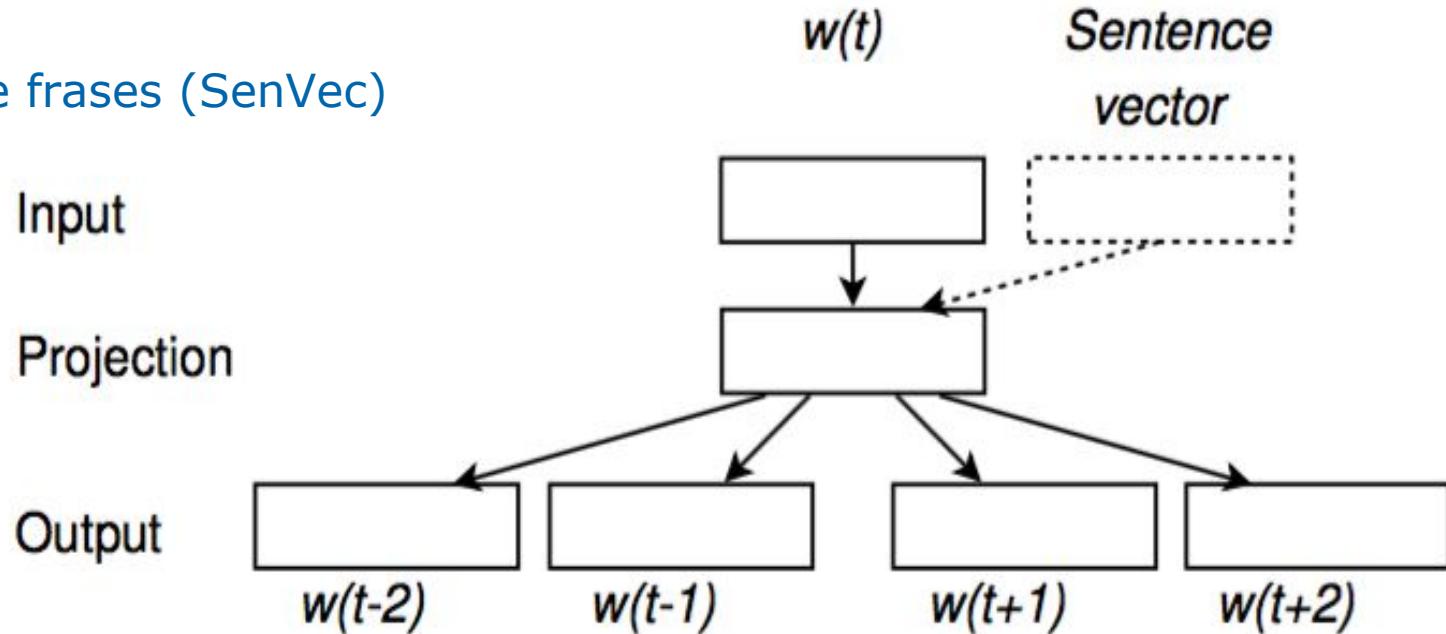
Representaciones del estado del arte

- Modelos basados en *n*-gramas.
- Iteramos *n* entre 1 y 10.
- Seleccionamos los 1.000, 5.000 y 10.000 *n*-gramas más frecuentes (o de mayor peso).
- Los mejores resultados se obtienen con:
 - 4-gramas de caracteres; los 10.000 más frecuentes.
 - 1-gramas de palabras (bag-of-words); los 10.000 más frecuentes.
 - 2-gramas de palabras; los 10.000 con mayor tf-idf.

Representaciones distribuidas

Dos variantes del modelo skip-gramas continuo de Mikolov et al.:

- Skip-gram
- Vectores de frases (SenVec)



(Franco et al., CLEF 2015)

Identificación de variedades

Representación	Accuracy (%)
Skip-gram	72,2*
LDR	71,1
SenVec	70,8**
BOW	52,7
Char 4-grams	51,5
EmoGraph	39,3
tf-idf 2-grams	32,2
Baseline aleatoria	20,0

* $z_{0.05} = 0,5457 < 1,960$

** $z_{0.05} = 0,7095 < 1,960$

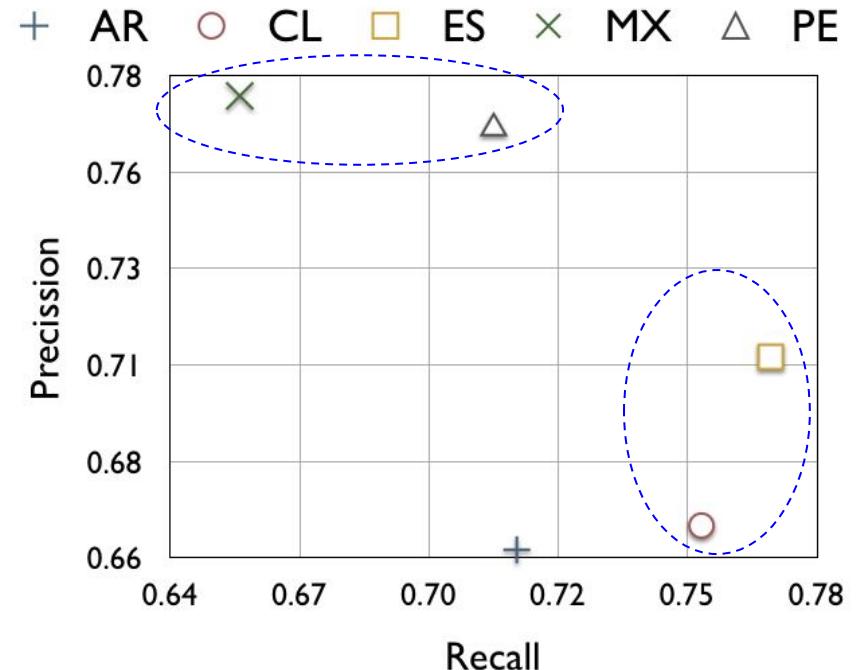
Análisis del error

Variety	Clasified as				
	AR	CL	ES	MX	PE
AR	143	16	22	8	11
CL	17	151	11	11	10
ES	20	13	154	7	6
MX	20	18	18	131	13
PE	16	28	12	12	132

Matriz de confusión en la clasificación a 5 clases

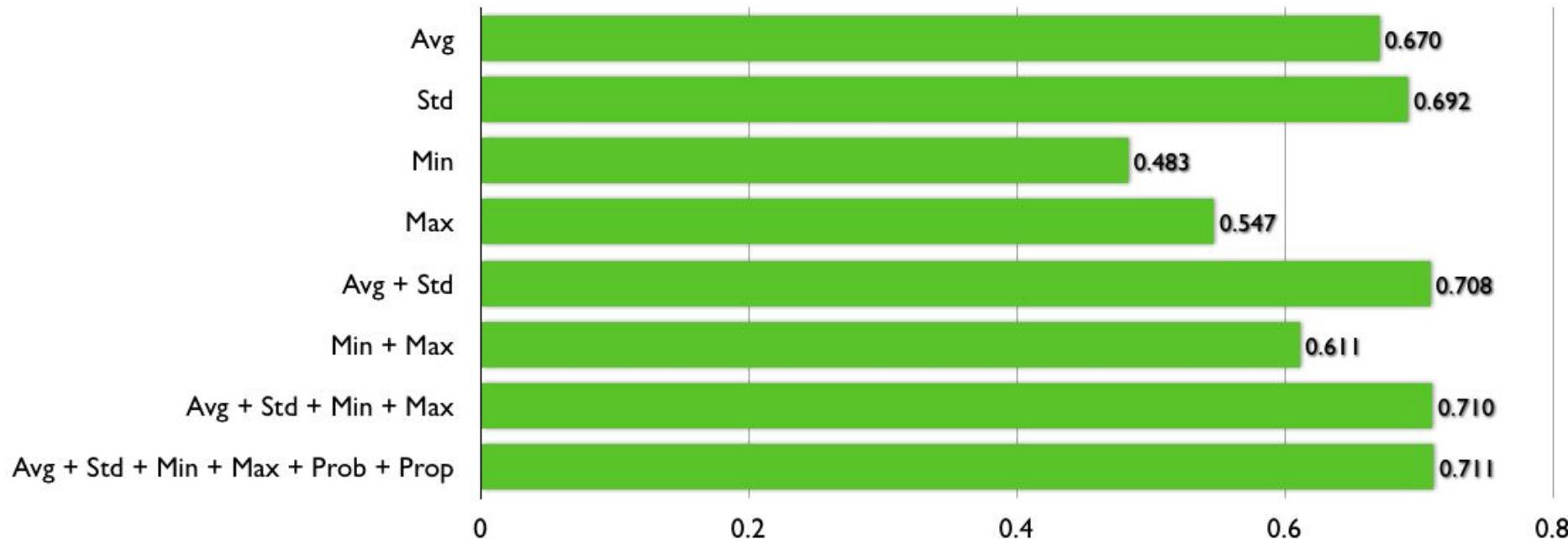
Introducción
Sexo y Edad
Variedad del Lenguaje
Conclusiones

Introducción
Low Dimensionality Representation (LDR)
Metodología
Resultados
Análisis y Discusión
Conclusiones



F1 para la identificación de la variedad correspondiente vs. las otras

Características más discriminantes



Accuracy obtenida con diferentes combinaciones de características

Análisis de costes

Introducción	Introducción
Sexo y Edad	Low Dimensionality Representation (LDR)
Variedad del Lenguaje	Metodología
Conclusiones	Resultados
	Análisis y Discusión
	Conclusiones

Complejidad de obtener las características:

$$O(l \cdot n) + O(l \cdot m) = O(\max(l \cdot n, l \cdot m)) = O(l \cdot n)$$

{ I: número de variedades
n: número de términos del documento
m: número de términos en el documento que coinciden con algún término del vocabulario
n \geq m & I << n

Número de características:

Representation	# Features
LDR	30
Skip-gram	300
SenVec	300
EmoGraph	1.011
BOW	10,000
Char 4-grams	10,000
tf-idf 2-grams	10,000

Robustez frente a lenguas

DSLCC

- Frases extraídas de noticias.
- Frases de entre 20 y 100 tokens.
- Número de instancias por conjunto:

Entrenamiento	Desarrollo	Pruebas
252.000	28.000	14.000

(Zampieri et al., DSL 2014)

Grupo	Lengua	Código
Eslavo sudoriental	Búlgaro Macedonio	bg mk
Eslavo sudoccidental	Bosnio Croata Serbio	bs hr sr
Eslavo occidental	Checo Eslovaco	cz sk
Español	Argentino Peninsular	es-AR es-ES
Portugués	Brasileño Europeo	pt-BR pt-PT
Austranesio	Indonesio Malayo	id my
Otros		xx

(Franco et al., DSL 2015; Fabra et al., DSL 2015)

Robustez frente a lenguas

Lengua	LDR	Skip-gram	SenVec	Lengua	LDR	Skip-gram	SenVec
Búlgaro	99,9	100	100	Bosnio	78,0*	80,3	74,4
Macedonio	99,9	100	100	Croata	85,8	85,9	84,7
Español-España	84,7*	82,1	86,3*	Serbio	86,4*	75,1	91,2
Español-Argentina	88,0	90,3*	87,6	Indonesio	99,4	99,3	99,4
Portugués-Portugal	87,4*	83,2	90,0*	Malayo	99,2	99,2	99,8*
Portugués-Brasil	90,0*	94,5*	87,6	Checo	99,8	99,9	99,8
Otras lenguas	99,9	99,8	99,8	Eslovaco	99,3	100*	99,3

*Resultados significativos con respecto al siguiente resultado

Identificación de sexo y edad

Dataset	Género	Idioma	Edad		Pos.	Sexo		Pos.	Nº Particip.
			EmoGraph	LDR		EmoGraph	LDR		
PAN-AP-2013	Social Media	Español	66,24*	62,70	3	63,65*	60,75	6	21
PAN-AP-2014	Social Media	Español	45,9	38,16	6	68,6*	56,89	9	9
PAN-AP-2014	Social Media	Inglés	34,2*	31,63	6	53,4	51,42	9	10
PAN-AP-2014	Blogs	Español	46,4	46,43	3	64,3	50,00	5	9
PAN-AP-2014	Blogs	Inglés	46,2	38,46	3	71,3	67,95	1	10
PAN-AP-2014	Twitter	Español	58,9	56,67	2	73,3	63,33	2	8
PAN-AP-2014	Twitter	Inglés	45,5	52,60	1	72,1	67,53	3	9
PAN-AP-2014	Revisiones	Inglés	30,8	32,28	5	66,1	67,11	5	10

- **LDR supera** a las representaciones comúnmente utilizadas en el estado del arte (n-gramas) por un **35%** de incremento en accuracy.
- LDR obtiene **resultados competitivos** en comparación a las representaciones distribuidas que emplean el popular y de alto rendimiento **modelo de Skip-grams continuo**.
- LDR se mantiene competitivo en diferentes **lenguas** y **medios** (DSLCC).
- La **reducción de la dimensionalidad** es de miles de características a sólo 6 características por variedad, lo que la hace idónea para tratar con **big data** en **social media**.
- Hemos aplicado LDR a la tarea de **identificación de edad y sexo**, obteniendo resultados competitivos.

- **EmoGraph** (grafos + emociones) permite corroborar nuestra hipótesis:

discurso + emociones --> edad y sexo

- Independiente del **medio social** y del **idioma**.
- Su contribución destaca cuando dispone de un **mínimo de palabras**.
- Su **coste computacional** permite su aplicación a entornos **big data** como los **medios sociales**.

- EmoGraph no es adecuado para la identificación de variedades del lenguaje:
 - El modo de organizar el discurso o la emotividad no varía entre variedades.
 - La variación se encuentra a nivel de contenido (e.g. vocabulario, expresiones...).
- **LDR** permite:
 - Obtener resultados **superiores** a modelos de **n-gramas** y **competitivos** con las **representaciones distribuidas**.
 - Reducir considerablemente la **dimensionalidad** y hacerlos óptimos para aplicar en **big data** de los **medios sociales**.

EmoGraph	LDR
<ul style="list-style-type: none">• Detección de bots• Perfiles políticos• Detección de engaño	<ul style="list-style-type: none">• HispaTweets• Variedades dialectales
<ul style="list-style-type: none">• Tendencia perceptiva y de aprendizaje (visuales, auditivas, kinestésicas)• Perfil motivacional (logro, reconocimiento, poder)	

- **18 publicaciones científicas**, a destacar:
 - A Low Dimensionality Representation to Language Variety Identification. In: Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (**CICLing'16**), Springer-Verlag, LNCS(), pp. , 2016
 - On the Impact of Emotions on Author Profiling. In: Information Processing & Management (**IP&M'16**) 52(1): 73-92, 2016
 - Emotional Trends in Social Media - A State Space Approach. In: 21st. Conference on Artificial Intelligence (**ECAI'14**) pp. 1123-1124, 2014
- Impacto en medios:
 - ¿Es Internet un Cerebro? - Informe Semanal **TVE**
 - Entrevista **UPV-Televisión** - impacto en **9 diarios digitales + 16 impresos**
 - ¿Está sobrevalorado el poder de los “influencers” en Twitter? - Diario **Expansión**
- Difusión en medios sociales (**+50.000 lectores**):
 - Socialancer, Autoritas Coolhunting, Hablemos de I+D



UNIVERSITAT POLITÈCNICA DE VALÈNCIA

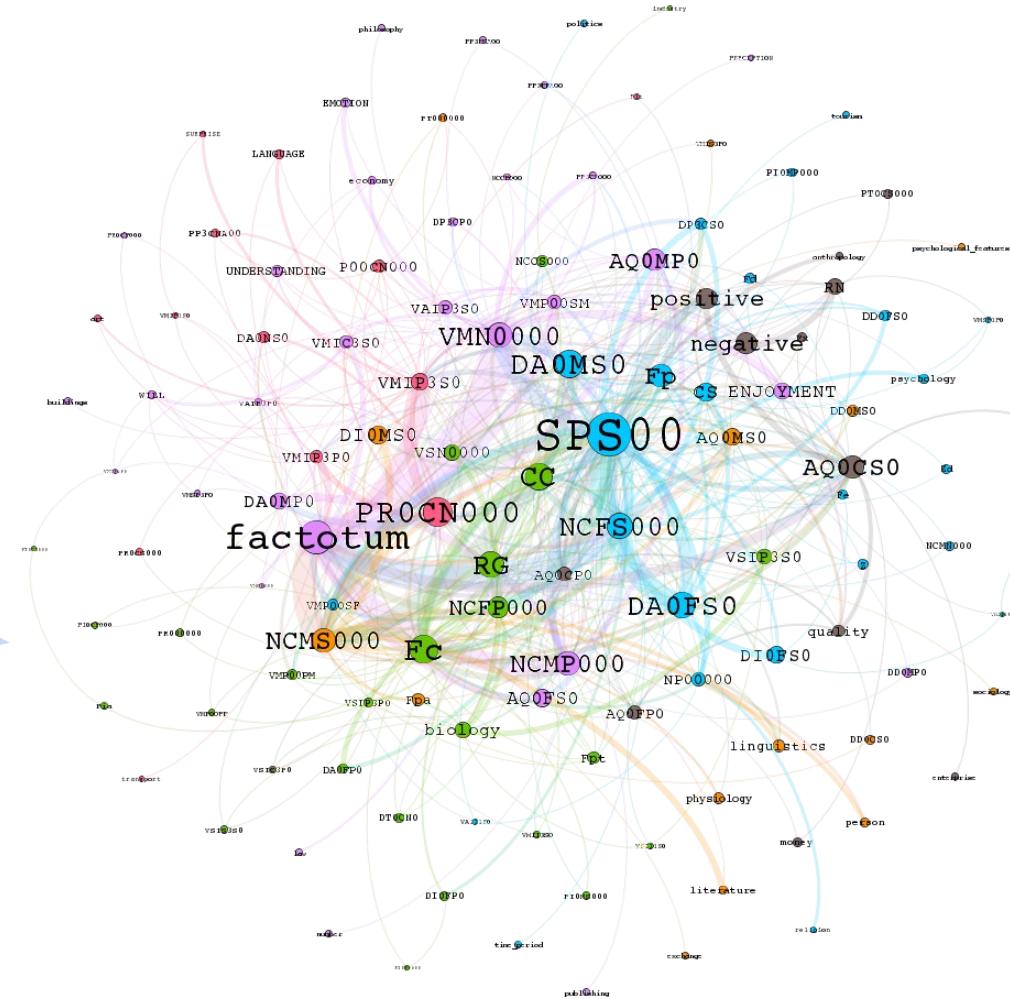
TESIS DOCTORAL

Author Profiling en Social Media: Identificación de Edad, Sexo y Variedad del Lenguaje

Autor:

Director:

Write to me...



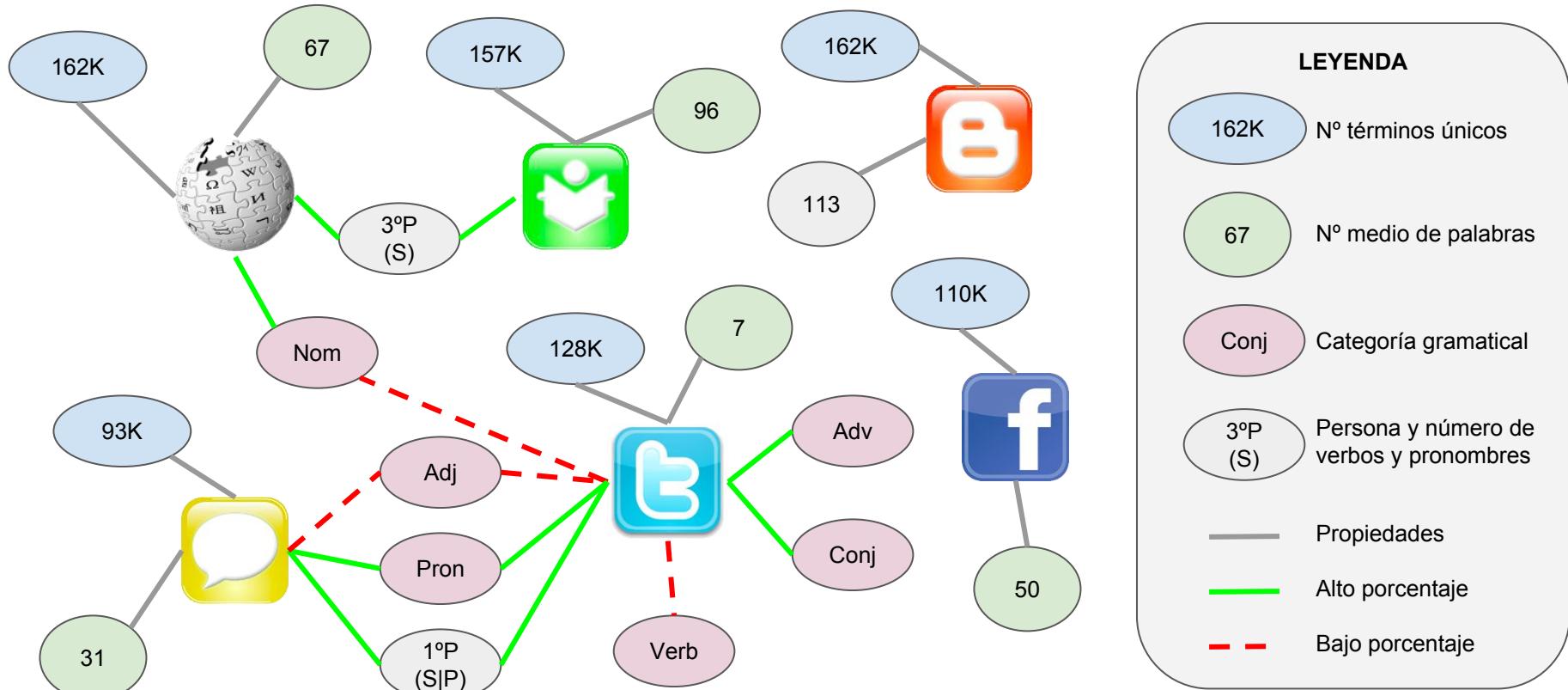
...and I'll profile you!



@kicorangel

MATERIAL ADICIONAL

% de uso de personas y números (Rangel & Rosso, Comunica 2013)



¿Define el canal el lenguaje que usamos? (Rangel & Rosso, Comunica 2013)

	Wikipedia	Prensa	Blogs	Foros	Twitter	Facebook
Documentos	3.987.179	5.191.694	1.083.709	673.664	23.873.371	576.723
Términos	267.465.810	499.477.658	122.509.753	21.026.388	163.188.488	28.947.716
Términos únicos	162.357	157.457	162.412	93.145	128.147	110.040
Ratio léxico*	1,89	1,83	1,89	1,08	1,49	1,28
Media palabras	67	96	113	31	7	50

*Con respecto a los 85.918 lemas de la 22^a edición del Diccionario de la Real Academia Española.

% de uso de categoría gramatical por medio (Rangel & Rosso, Comunica 2013)

	Wikipedia	Prensa	Blogs	Foros	Twitter	Facebook
Adjetivo	13,57	12,50	13,67	9,27	6,62	12,06
Adverbio	2,78	3,46	3,87	4,74	6,30	3,49
Conjunción	1,52	2,10	1,80	4,18	7,00	2,65
Cuantitativo	3,34	4,47	4,15	5,34	5,53	4,29
Determinante	2,88	3,48	2,78	4,18	6,40	4,02
Preposición	4,00	5,49	5,07	8,94	13,81	6,15
Pronombre	0,65	0,92	1,12	2,22	3,32	1,39
Nombre	50,33	47,05	46,59	42,63	34,08	47,04
Verbo	20,55	20,47	20,88	18,08	16,56	18,83

% de uso de personas y números (Rangel & Rosso, Comunica 2013)

	Persona	Número	Wikipedia	Prensa	Blogs	Foros	Twitter	Facebook
Pronombre	1	SIN	13,61	14,58	18,85	54,47	65,81	22,30
		PLU	0,00	0,00	0,00	0,00	0,00	0,00
	2	SIN	4,56	1,18	2,23	1,54	3,53	3,95
		PLU	1,92	1,75	5,31	4,61	5,62	3,49
	3	SIN	55,06	50,75	39,26	24,08	12,70	34,68
		PLU	13,42	18,22	16,93	8,91	3,35	17,14
	OTROS		11,41	13,52	17,42	6,39	8,99	18,44

% de uso de personas y números (Rangel & Rosso, Comunica 2013)

	Persona	Número	Wikipedia	Prensa	Blogs	Foros	Twitter	Facebook
Verbo	1	SIN	19,95	17,41	17,50	28,94	24,00	16,61
		PLU	2,10	2,42	4,19	2,68	4,68	4,89
	2	SIN	6,02	1,55	3,58	3,55	6,77	2,95
		PLU	0,46	0,42	0,69	0,98	1,65	0,76
	3	SIN	31,40	34,00	29,92	28,80	31,21	31,21
		PLU	40,07	44,20	45,11	35,05	31,69	43,59

Palabras más frecuentes (Rangel & Rosso, Comunica 2013)

Wikipedia	Prensa	Blogs	Foros	Twitter	Facebook
de	de	a	de	de	de
en	la	de	y	que	la
la	el	la	que	a	el
y	en	en	a	la	en
el	a	el	la	el	y
por	que	y	el	y	a
un	y	que	en	en	que
una	del	del	un	no	los
que	los	los	no	me	del
a	por	un	pregunta	un	por
los	un	por	es	es	para
del	se	se	por	se	un
es	con	con	abierta	lo	con
las	las	para	se	con	se
con	para	las	para	por	no

Emociones y Tendencias en Twitter (Volgmann et al., ECAI 2014)

Motivación

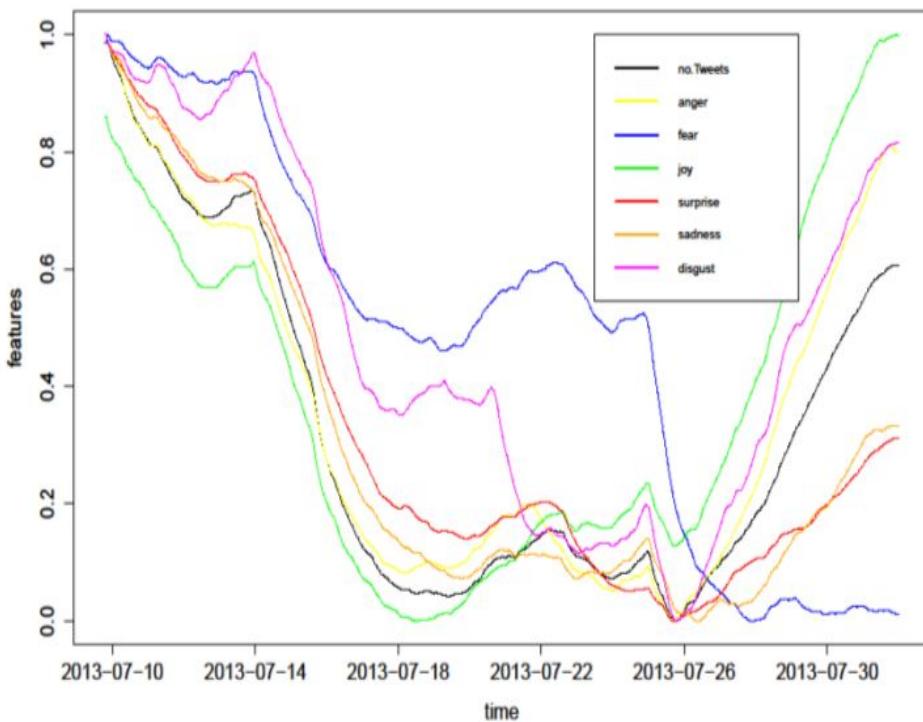
- El número de tuits en un momento depende de la dinámica previa de la discusión.
- Los usuarios responden a estímulos emocionales.

Aproximación

- Conversación en Twitter sobre el caso Bárcenas (julio-octubre 2013; 4,3M de tuits).
- Modelo de espacio de estados con las 6 emociones de Ekman.
- Estimación de tendencias mediante filtrado Kalman.

Conclusiones

- El sistema deduce la evolución de la conversación a partir de las emociones, haciendo los cambios más aparentes.
- El modelo de espacio de estados es eficiente con ciclos y ruido.

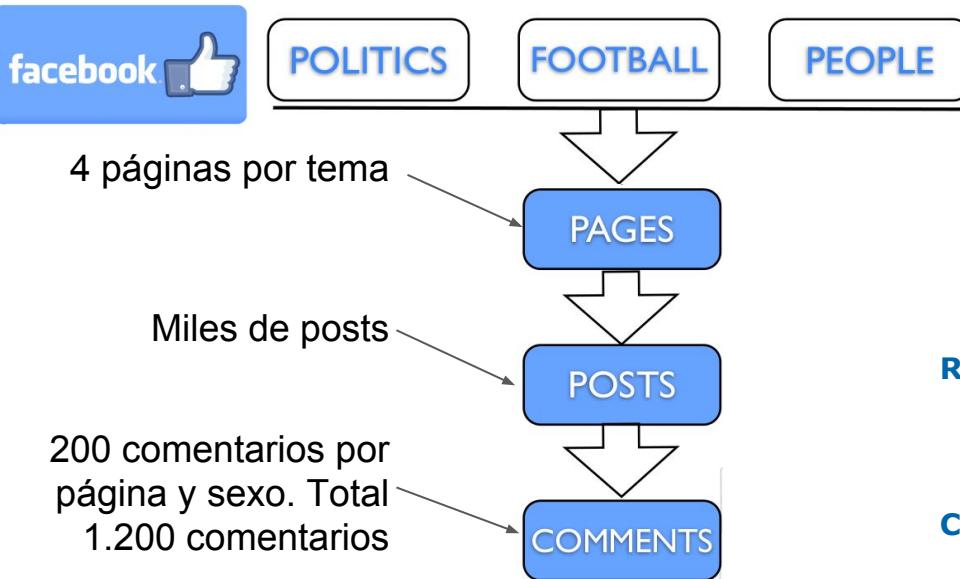


Características estilo + emociones (Rangel & Rosso, NLPCS 2013)

Frecuencias	Ratio entre el número de palabras únicas y el número total de palabras (unique-words), palabras comenzando por mayúsculas (capital-words), palabras en mayúsculas (upper-words), longitud de las palabras (words-length), número de letras en mayúsculas (upper-chars) y número de palabras con character flooding (e.g. Hoooooolaaaaaa) (running-words).
Signos de puntuación	Frecuencia de uso de signos de puntuación (punctuation), de puntos (punctuation-dots), comas (punctuation-commas), dos puntos (punctuation-colon), punto y coma (punctuation-semicolo), exclamaciones (punctuation-exclamation), preguntas (punctuation-question) y comillas (punctuation-quotes).
Categorías gramaticales	Frecuencia de uso de cada categoría gramatical (Adj., Adv., Conj., Quant., Det., Intj., MD., Prep., Pron., Noun., Verb.), número y persona de verbos (Verb1S, Verb2S, Verb3S, Verb1P, Verb2P, Verb3P) y pronombres (Pron1S, Pron2S, Pron3S, Pron1P, Pron2P, Pron3P), modo de los verbos (VerbIndicative, VerbSubjunctive, VerbConditional, VerbImperative, VerbInfinitive, VerbParticiple), nombres propios (Propper) y palabras no reconocidas (NotRecognized).
Emoticonos	Ratio entre el número de emoticonos y el número total de palabras (emoticons), número de tipos diferentes de emoticonos: alegría (emoticon-happy), tristeza (emoticon-sad), disgusto (emoticon-disgust), enfado (emoticon-angry), sorpresa (emoticon-surprise), burla (emoticon-wink) o tontería (emoticon-dumb).
Lexicón de emociones en español (SEL*)	Para cada palabra se obtiene su lema. Para el lema, se busca en SEL su factor de probabilidad de uso afectivo. Si el lema no tiene una entrada en el diccionario, se obtienen sus sinónimos. Para cada sinónimo, se busca el citado factor. Para cada emoción, se suman todos los valores obtenidos, creando una característica por emoción (sel-enjoyment, sel-surprise, sel-anger, sel-disgust, sel-sadness, sel-fear). <small>*Sidorov et al., 2012</small>

Emociones y Sexo en Facebook (Rangel & Rosso, ESSEM 2013)

Dataset - EmIroGeFB



Emociones

- Etiquetados con las 6 emociones básicas de Ekman.
- Tres anotadores. Concordancia: 14,55%

	MUJERES	HOMBRES
Determinante	6,81	7,74
Interjección	0,18	0,30
Preposición	6,25	5,85
Pronombres	2,24	2,67

Resultados

- $r=18$ y accuracy=59% - Valor comparativo al PAN.

Conclusiones

- Las características usadas para identificar emociones, permiten identificar el sexo, lo que sugiere cierta correlación entre el uso de emociones y el sexo del autor.

Emociones, Sexo e Ironía en Facebook (Rangel et al., LREC 2014)

Motivación

- Relación entre ironía, emociones y sexo.

Aproximación

- Etiquetado manual de la ironía en EmIroGeFB.
- Tres anotadores. Concordancia de -6,60%.
- "Pitbul es cultura, ¿no ves que te enseña a contar? aunque sea sólo hasta 3" -> 2 de 3 anotaron ironía

Conclusiones

- Las mujeres tienden a usar más emociones que los hombres.
- Los hombres tienden a ser más irónicos.
- En política se expresan más emociones negativas e ironía.

EMOC.	TOTAL	Mujeres	Hombres	Política	Fútbol	Famoseo
Alegría	338 / 8	194	144	50	153	135
Disgusto	129 / 6	63	66	79	7	43
Enfado	151 / 4	72	79	114	10	27
Miedo	3 / 0	1	2	2	1	0
Sorpresa	390 / 6	215	175	53	180	157
Tristeza	76 / 0	39	37	52	9	15
Ninguna	262 / 3	18	37	9	23	23

IRÓNICOS	Política	Fútbol	Famoseo	TOTAL
Mujeres	11	1	3	15
Hombres	16	3	8	27
TOTAL	27	4	12	42

Emociones, Sexo y Edad en PAN (Rangel & Rosso, NLPCS 2013)

Dataset - PAN-AP2013

- Equilibrado por sexo.
- Edades: 10s (13-17), 20s (23-27), 30s (33-47).

Edad	Nº de Autores	
	Entrenamiento	Pruebas
10s	2.500	240
20s	42.600	3.840
30s	30.800	2.720

Aproximación

- Características de estilo.
- Máquinas de vectores soporte. Kernel Gausiano con $g=0.01$ y $c=2000$ (Weka).
- Evaluación mediante Accuracy

Resultados

TAREA	SEXO	EDAD
Posición	7 / 22	3 / 22
Accuracy	57,13%	63,50%
Acc. mejor equipo	64,73%	65,58%

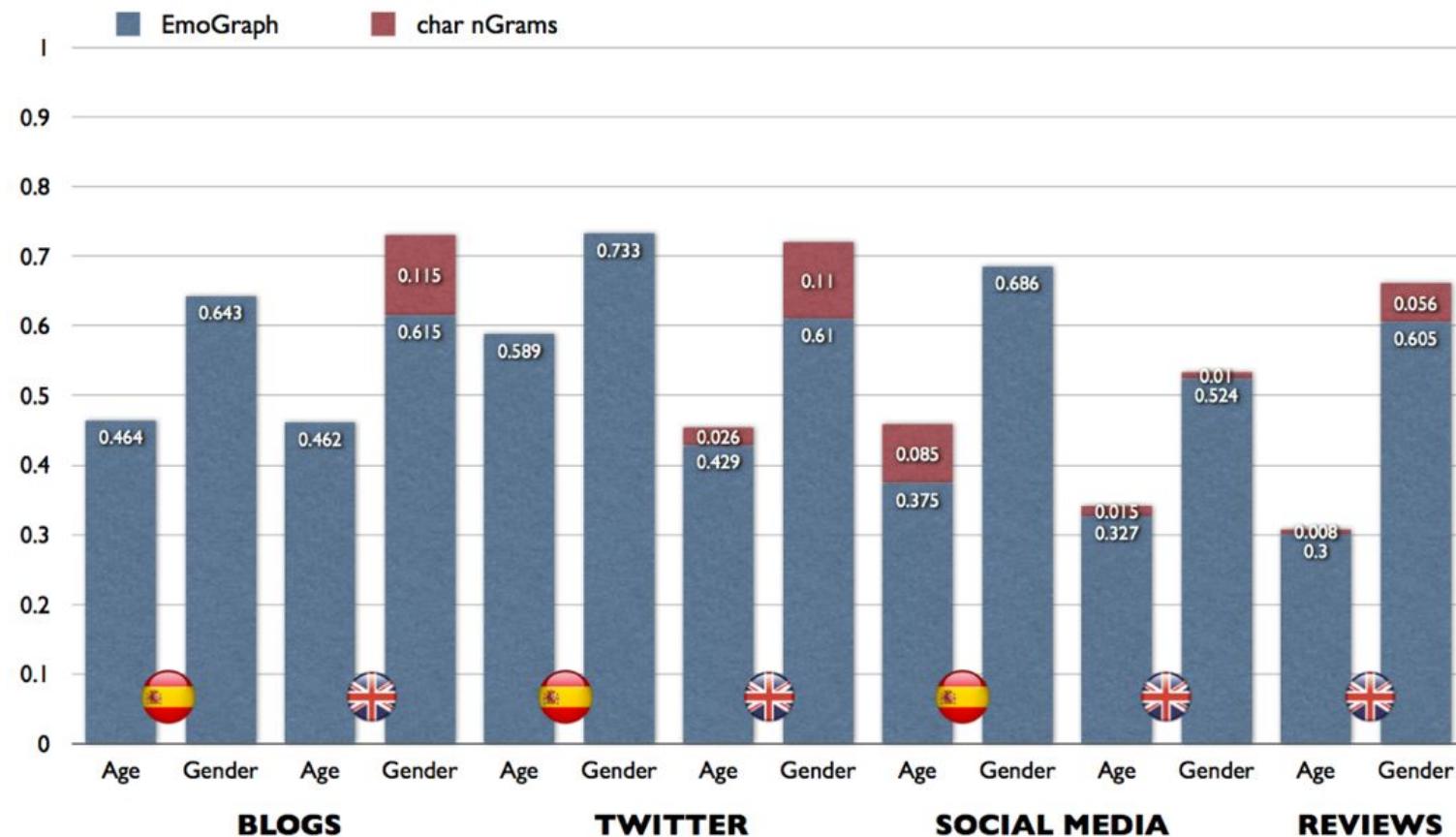
Conclusiones

- Las características de estilo obtienen resultados competitivos, especialmente en identificación de edad.
- De nuevo se sugiere la relación entre la expresión de emociones y la edad y el sexo del autor.

Recursos EmoGraph inglés (Rangel & Rosso, CLEF 2015)

Freeling	http://nlp.lsi.upc.edu/freeling
WordNet Domains	http://wndomains.fbk.eu
Clasificación semántica de verbos	Levin, B. English Verb Classes and Alternations. University of Chicago Press, Chicago. (1993)
Lexicón de polaridad	Hu, M., Liu, B. Mining and Summarizing Customer Reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Seattle, Washington, USA, pp. 168-177 (2004)
Wordnet Affect	C. Strapparava and A. Valitutti. <i>Wordnet-affect: an affective extension of wordnet</i> . In Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, 2004.

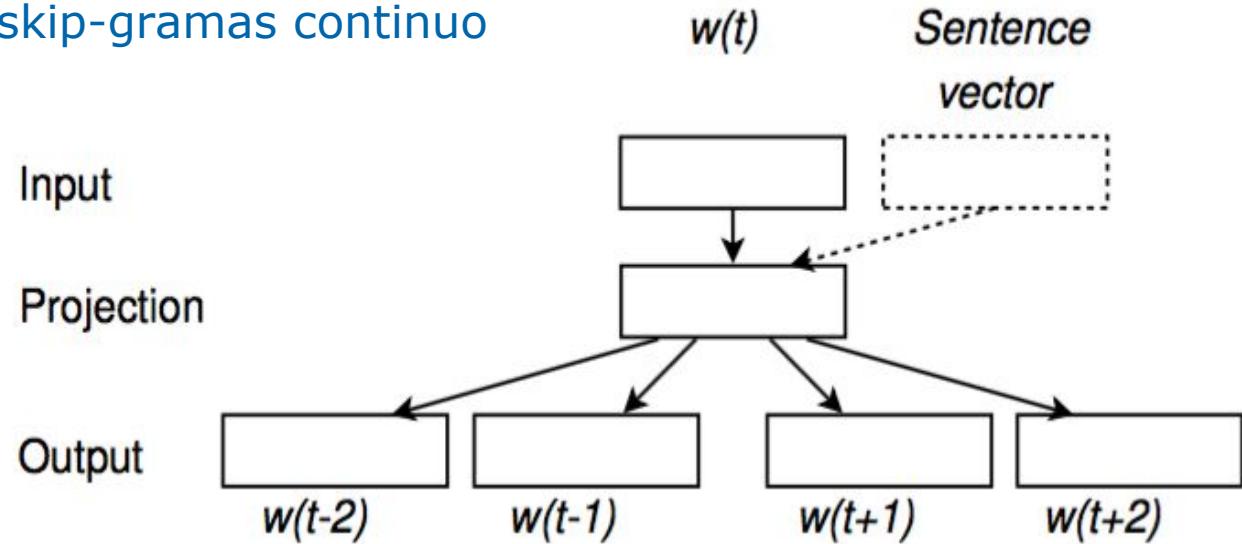
Contribución de EmoGraph en PAN-AP14 (Rangel et al., CLEF 2015)



Representaciones distribuidas (Franco et al., CLEF 2015)

Dos variantes del modelo skip-gramas continuo de Mikolov et al.:

- Skip-grams
- Vectores de frases (SenVec)



Maximizando la media de la función de log probabilidad: Usando el estimador de ejemplos negativos:

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

$$\log \sigma(v'_{w_O} {}^T v_{w_I}) + \sum_{i=1}^k \mathbb{E}_{w_i} \sim P_n(w) \left[\log \sigma(-v'_{w_i} {}^T v_{w_I}) \right]$$

(Franco et al., CLEF 2015)

Information Gain Word Patterns (Franco et al., CLEF 2015)

- El objetivo es obtener patrones léxico-semánticos para representar el contenido de los documentos. El método se basa en la hipótesis de construcción de patrones (REF):

“those contexts that are relevant to the definition of a cluster of semantically related words tend to be (part of) lexico-syntactic constructions”

- Estructura: patrón = (lema_u , dirección dependencia, etiqueta dependencia, lema_v)
 $(\text{bigote}_{c_{25}} \text{ (moustache)}, <, \text{dobj}, \text{afeitar}_{c_{643}} \text{ (to shave)}),$
 $(\text{peluca}_{c_{25}} \text{ (wig)}, <, \text{dobj}, \text{peinar}_{c_{643}} \text{ (to comb)}),$
 $(\text{pelo}_{c_{25}} \text{ (hair)}, <, \text{subj}, \text{encrespar}_{c_{643}} \text{ (to curl)})$
- En los experimentos hemos seleccionado como características las 1.000 palabras más frecuentes obtenidas de los patrones con mayor ganancia de información.

(Franco et al., 2015)

Algoritmos de aprendizaje

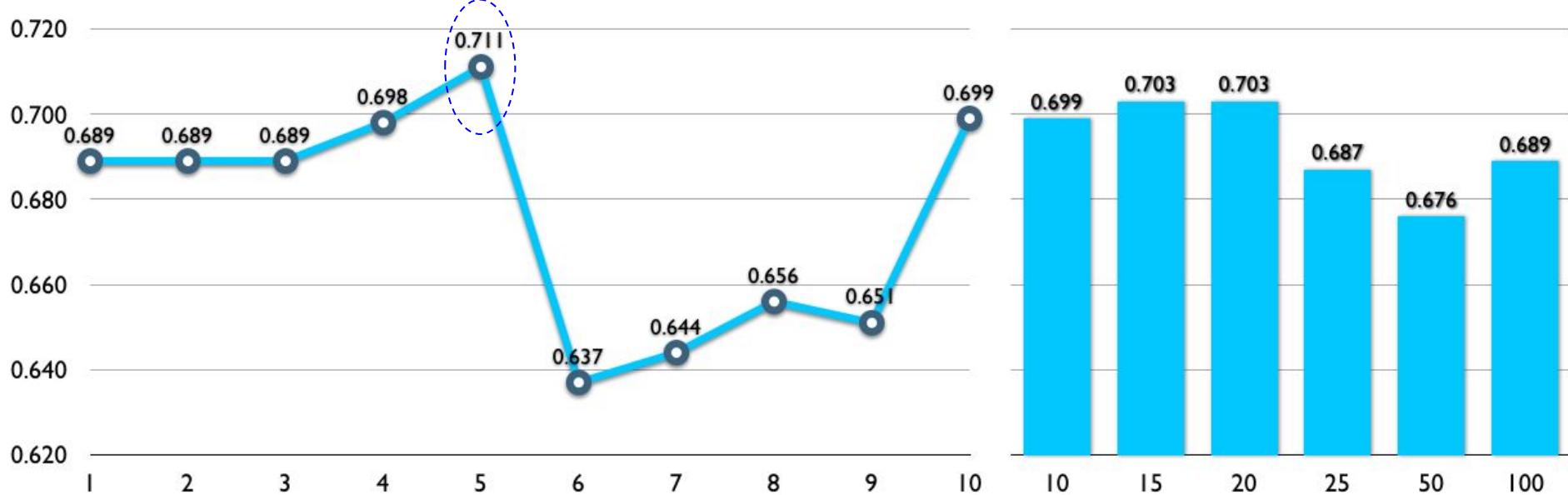
Algorithm	Accuracy	Algorithm	Accuracy	Algorithm	Accuracy
Multiclass Classifier	71.1	Rotation Forest	66.6	Multilayer Perceptron	62.5
SVM	69.3	Bagging	66.5	Simple Cart	61.9
LogitBoost	67.0	Random Forest	66.1	J48	59.3
Simple Logistic	66.8	Naive Bayes	64.1	BayesNet	52.2

Resultados en accuracy de los diferentes algoritmos de aprendizaje

Significación de los resultados del mejor resultado con respecto a los dos siguientes:

SVM ($z_{0.05} = 0,880 < 1,960$)
LogitBoost ($z_{0.05} = 1,983 > 1,960$)

Impacto del preprocesamiento



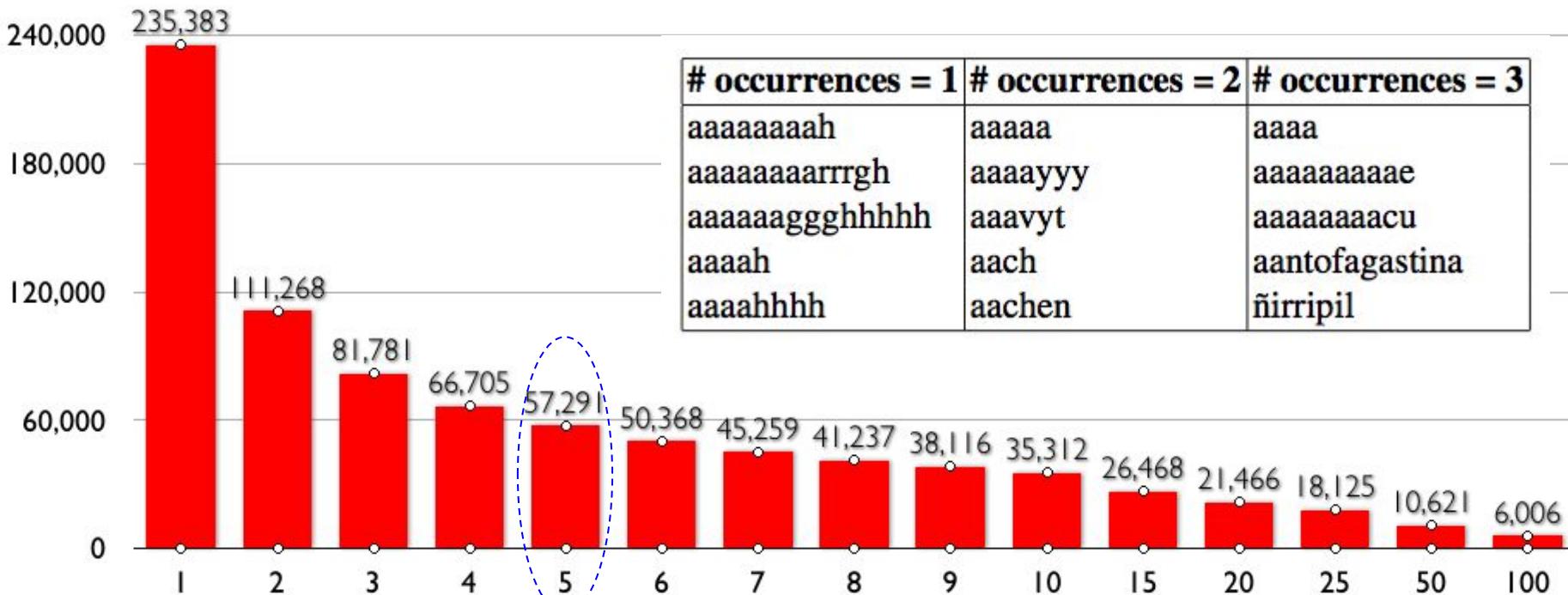
(a)

(b)

Accuracy tras de eliminar palabras con frecuencia igual o menor que n

(a) Escala continua (b) Escala no continua

Impacto del preprocesamiento



Número de palabras tras eliminar aquellas con frecuencia igual o menor que n , con algunos ejemplos de palabras infrecuentes.

Ganancia de información

Attribute	IG	Attribute	IG	Attribute	IG
PE-avg	0.680 ± 0.006	ES-std	0.497 ± 0.008	PE-prob	0.152 ± 0.005
AR-avg	0.675 ± 0.005	CL-max	0.496 ± 0.005	MX-prob	0.151 ± 0.005
MX-max	0.601 ± 0.005	CL-std	0.495 ± 0.007	ES-prob	0.130 ± 0.011
PE-max	0.600 ± 0.009	MX-std	0.493 ± 0.007	AR-prob	0.127 ± 0.006
ES-min	0.595 ± 0.033	CL-min	0.486 ± 0.013	AR-prop	0.116 ± 0.005
ES-avg	0.584 ± 0.004	AR-std	0.485 ± 0.005	MX-prop	0.113 ± 0.006
MX-avg	0.577 ± 0.008	PE-std	0.483 ± 0.012	PE-prop	0.112 ± 0.005
ES-max	0.564 ± 0.007	AR-min	0.463 ± 0.012	ES-prop	0.110 ± 0.007
AR-max	0.550 ± 0.007	CL-avg	0.455 ± 0.008	CL-prop	0.101 ± 0.005
MX-min	0.513 ± 0.027	PE-min	0.369 ± 0.019	CL-prob	0.087 ± 0.010

Visual, auditiva, kinestésica

Palabras relacionadas

V Ya veo, Observo, Imagino, Perspectiva

A Digo, Escucho, "En otras Palabras", Oye

K Siento, dame una mano, lo tengo, capto



Características del SDR visual

- Visualiza ayuda, identifica y establece ideas y conceptos.
- Aprende mejor cuando lee o visualiza la información.
- La abstracción y planificación se relaciona con la capacidad de visualizar.
- Al pensar a través de imágenes, se trae a la mente diversa información al mismo tiempo, igualmente al abstraer grandes cantidades de información.



Características del SDR auditivo

- Aprende mejor cuando recibe explicaciones orales y cuando se habla y explica la información a otra persona.
- Recuerda de manera ordenada y secuencial.
- El sistema auditivo no relaciona ni elabora conceptos abstractos con igual facilidad que el visual ni es tan rápido; sin embargo, es fundamental en el aprendizaje de idiomas y la música.
- Al memorizar de forma auditiva no olvidan ni una palabra.



Características del SDR kinestésico

- Presentan una distinta manera de aprender, menos rápida que los demás, ya que aprenden al hacer cosas.
- Se utiliza para aprender un deporte u otras actividades.
- Se procesa la información, asociándola a las propias sensaciones y movimientos del cuerpo.
- El aprendizaje kinestésico es profundo y duradero.

Logro, reconocimiento, poder



Temas por sexo (PAN-AP14 SM-EN)

Introducción ***Sexo y Edad*** Variedad del Lenguaje Conclusiones

Introducción
EmoGraph
Metodología
Resultados
Análisis y Discusión
Conclusiones

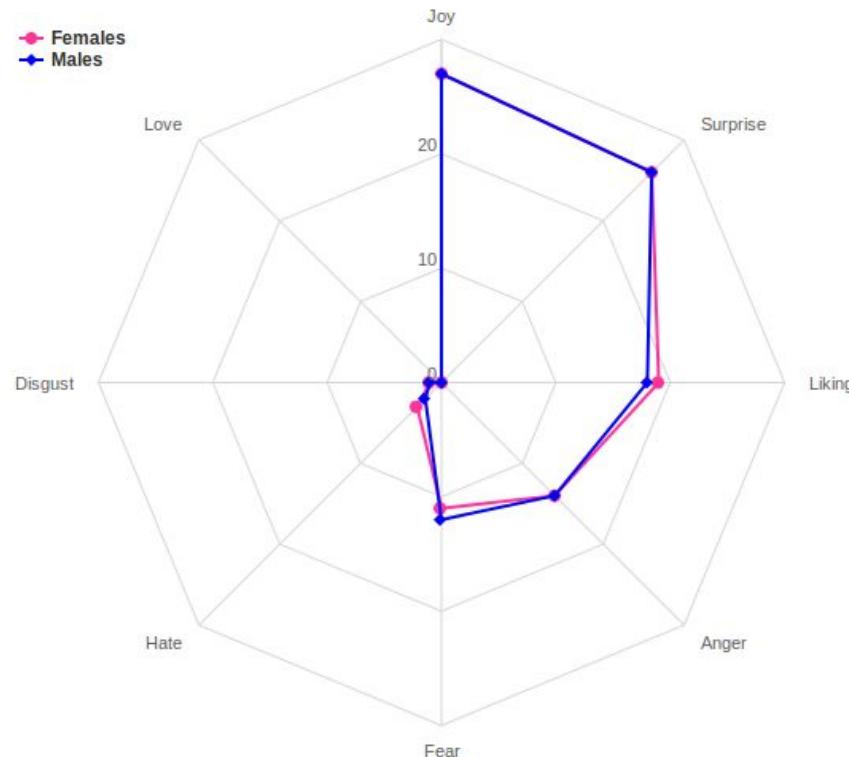
MUJERES

HOMBRES



- Sin diferencias significativas entre sexos.

Emociones por sexo (PAN-AP14 SM-EN)

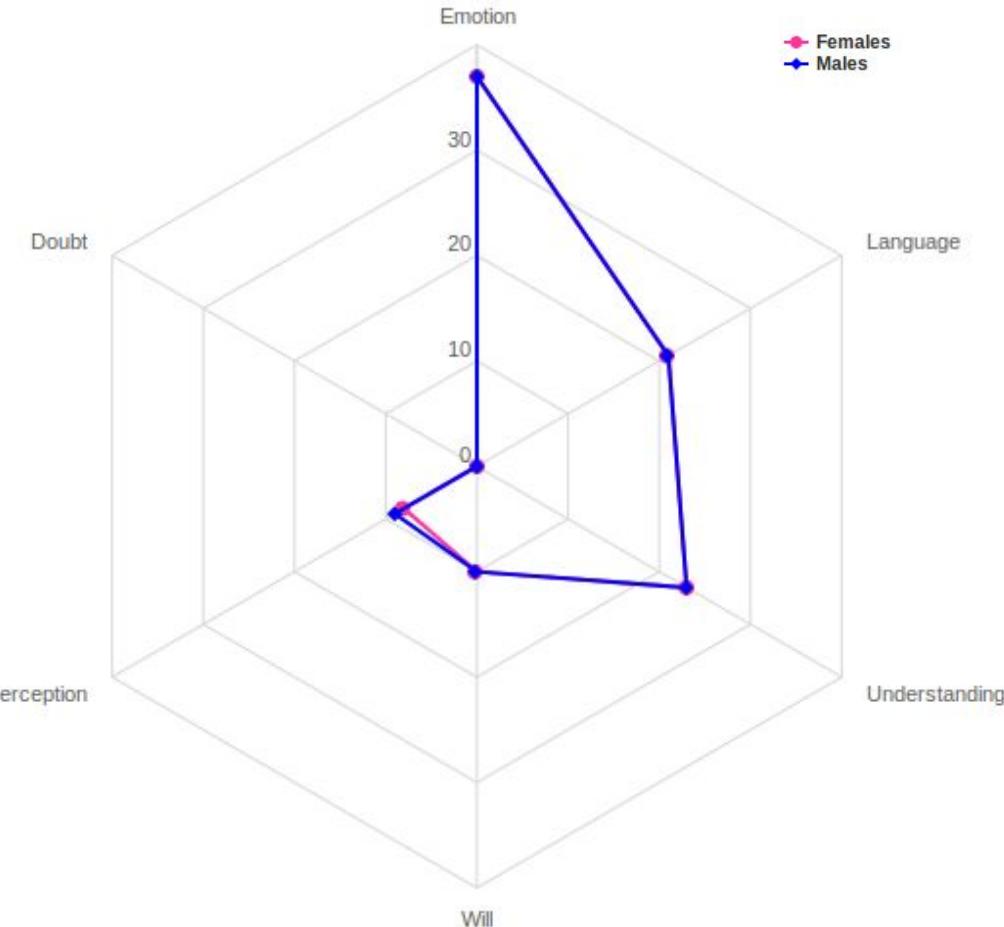


- Sin diferencias significativas entre sexos.

Tipos de verbos por sexo

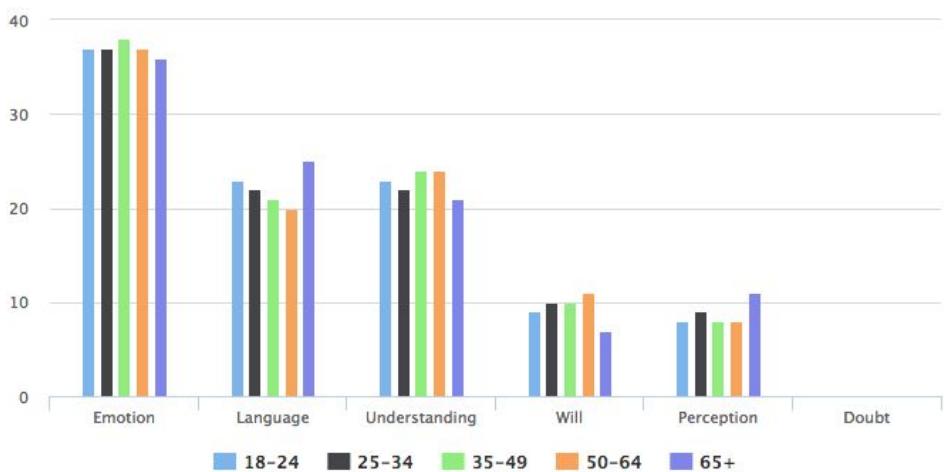
(PAN-AP14 SM-EN)

- Sin diferencias significativas entre sexos.

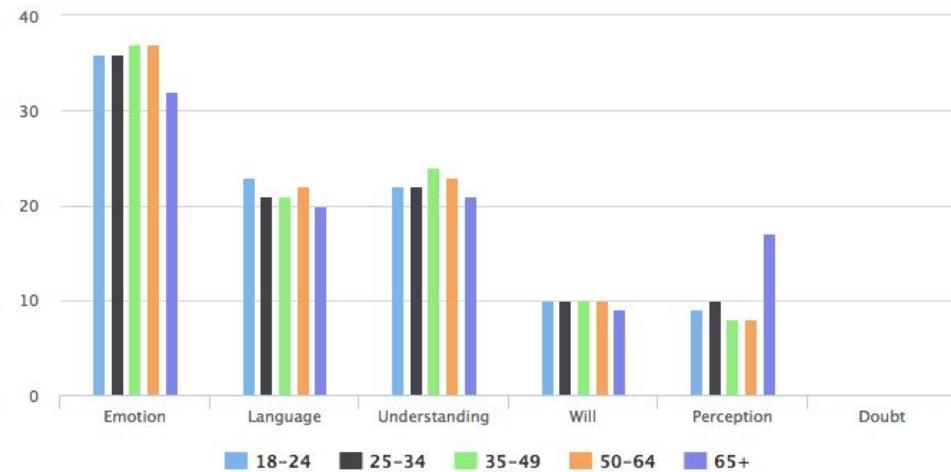


Evolución t. de verbos por sexo (PAN-AP14 SM-EN)

MUJERES



HOMBRES



Temas por sexo (PAN-AP14 SM-ES)

MUJERES

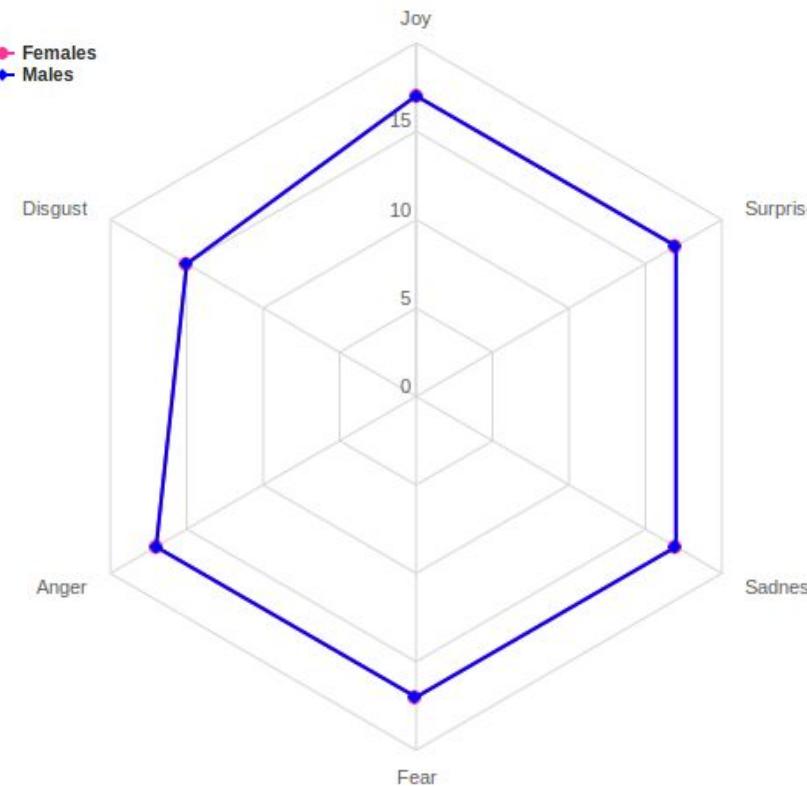


HOMBRES



- Sin diferencias significativas entre sexos.

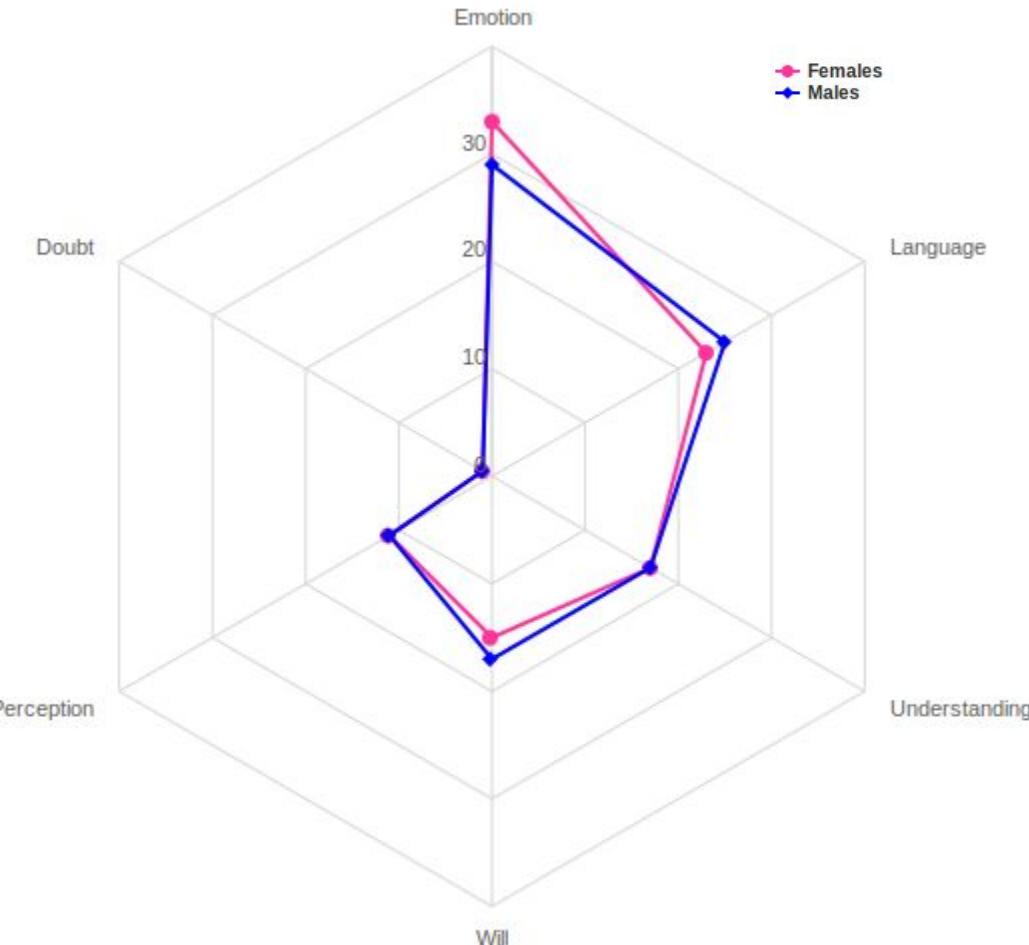
Emociones por sexo (PAN-AP14 SM-ES)



- Sin diferencias significativas entre sexos.

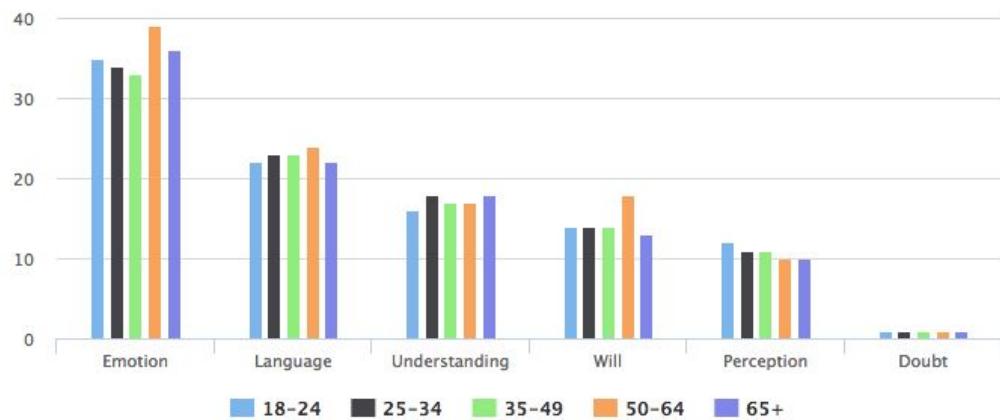
Tipos de verbos por sexo (PAN-AP14 SM-ES)

- Las mujeres usan más verbos de emoción (sentir, querer, amar...).
- Los hombres usan más verbos relativos al lenguaje (decir, contar, hablar...).

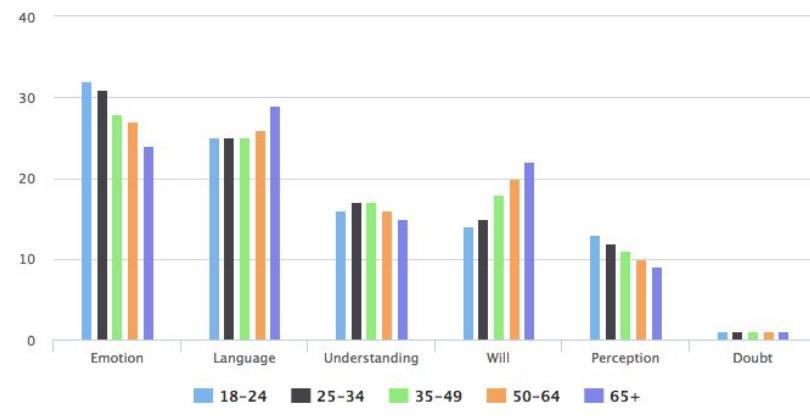


Evolución t. de verbos por sexo (PAN-AP14 SM-ES)

MUJERES

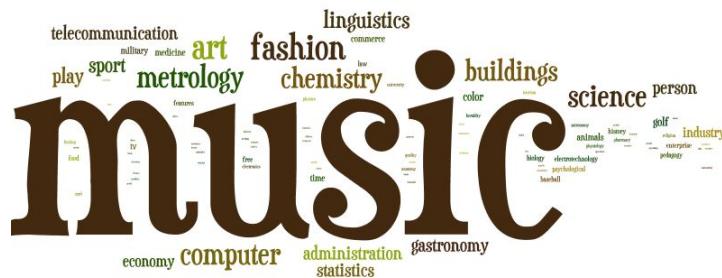


HOMBRES

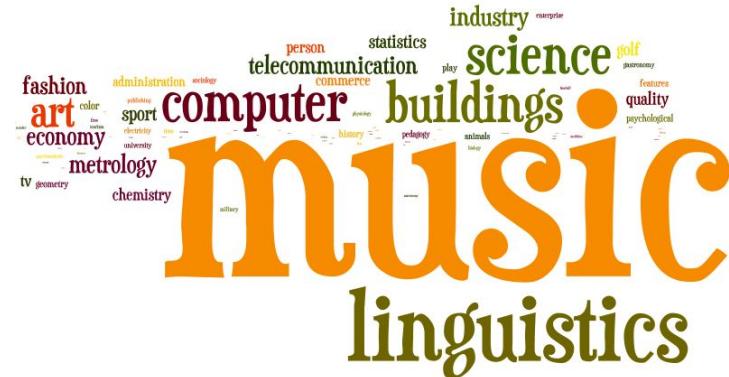


Temas por sexo (PAN-AP14 BL-EN)

MUJERES

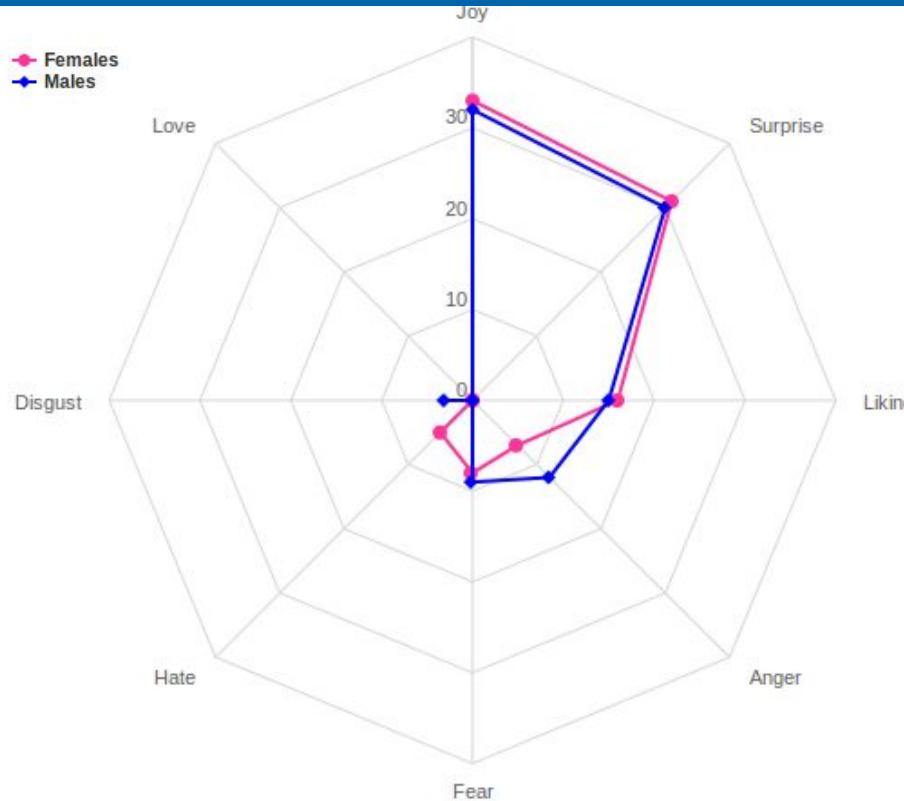


HOMBRES



- Aún sin excesivas diferencias, la mujer destaca por la lingüística o el arte frente al hombre.

Emociones por sexo (PAN-AP14 BL-EN)

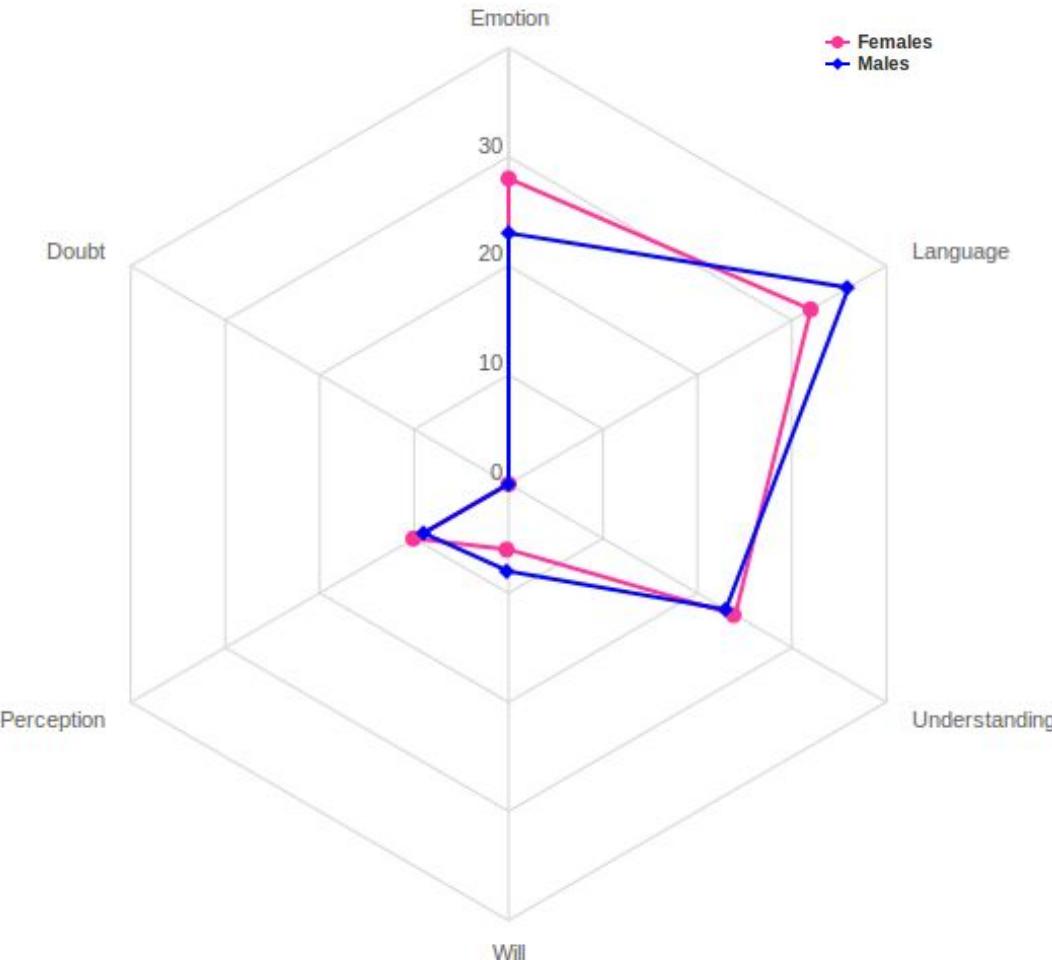


- Los hombres expresan más enfado y disgusto.
- Las mujeres expresan más odio.

Tipos de verbos por sexo

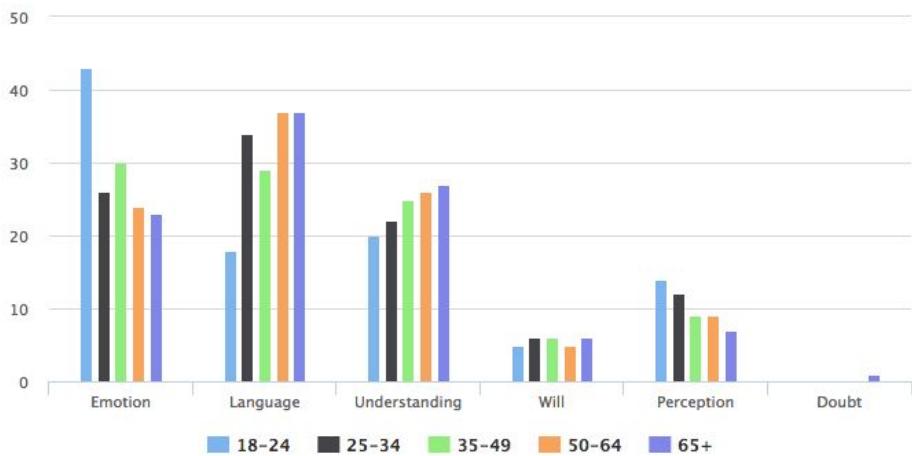
(PAN-AP14 BL-EN)

- Las mujeres usan más verbos de emoción (sentir, querer, amar...).
- Los hombres usan más verbos relativos al lenguaje (decir, contar, hablar...).

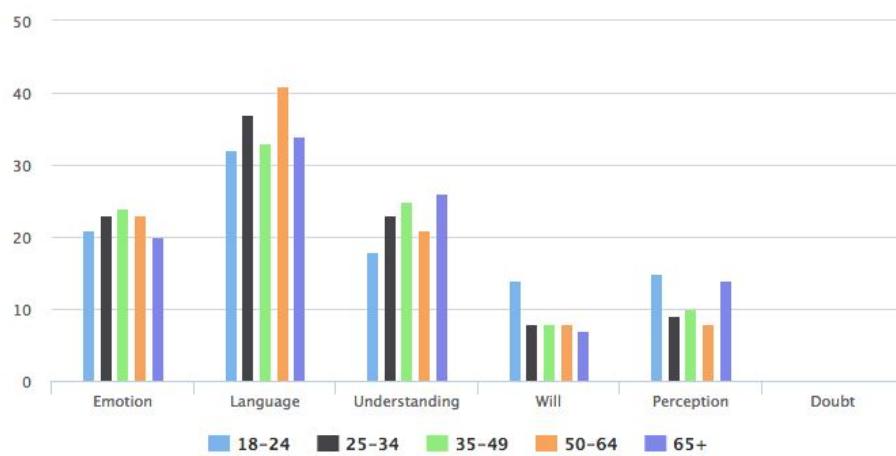


Evolución t. de verbos por sexo (PAN-AP14 BL-EN)

MUJERES

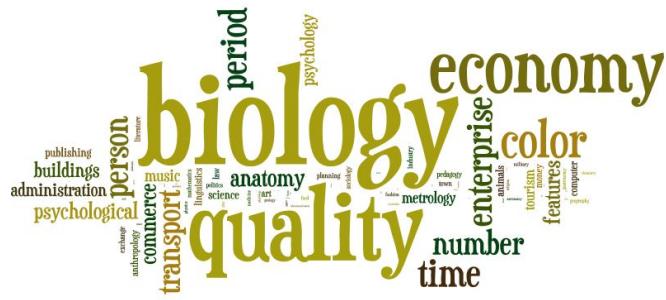


HOMBRES



Temas por sexo (PAN-AP14 BL-ES)

MUJERES

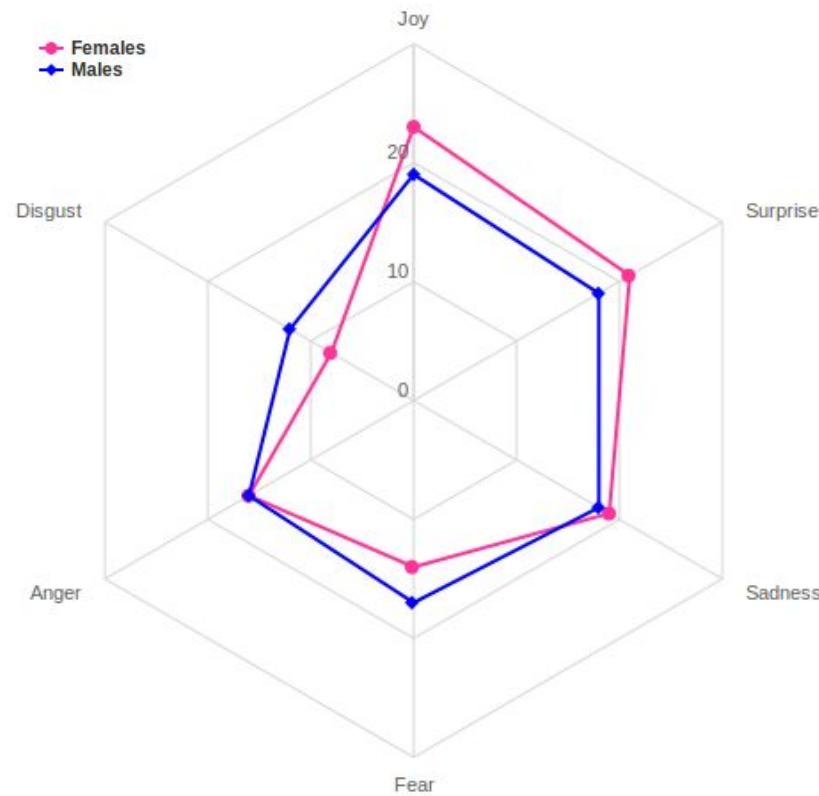


HOMBRES



- Sin diferencias significativas entre sexos.

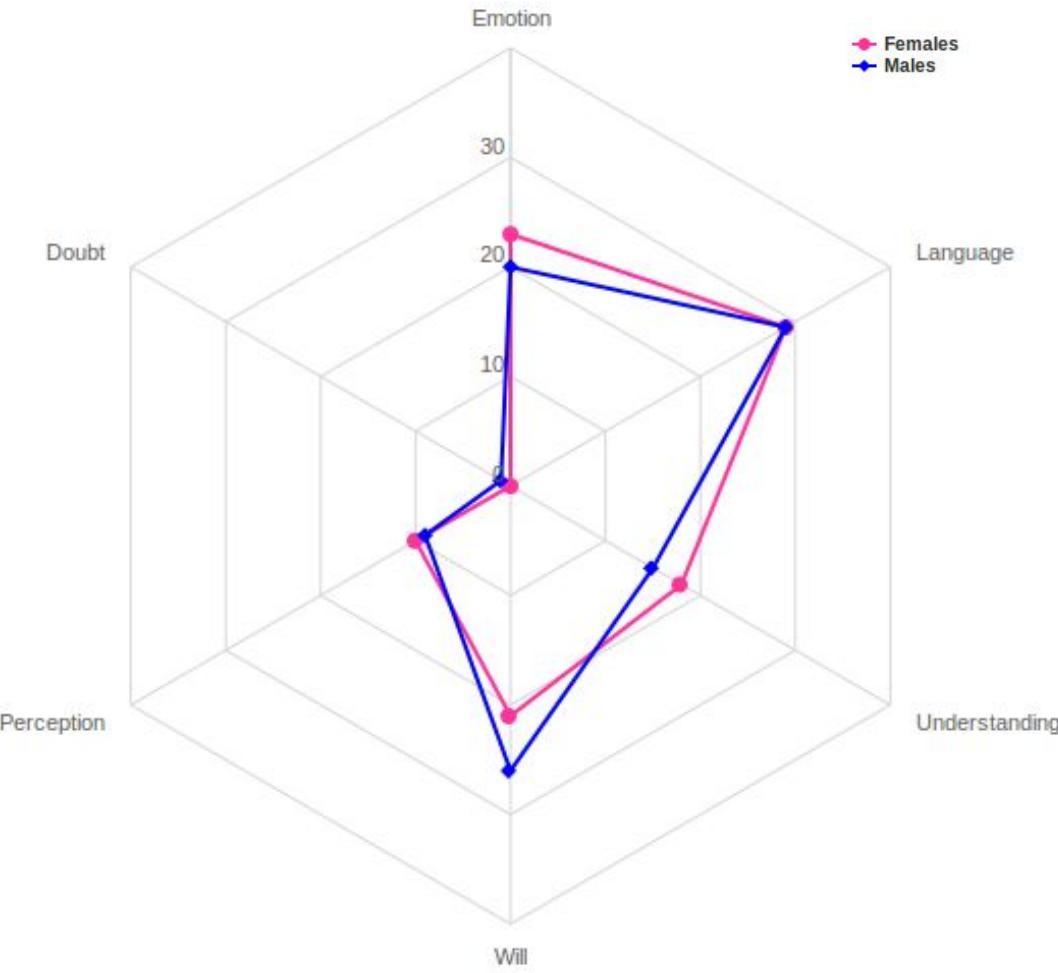
Emociones por sexo (PAN-AP14 BL-ES)



- › Las mujeres parecen expresar más alegría, sorpresa y tristeza.
- › Los hombres por su parte lo hacen con disgusto y miedo.

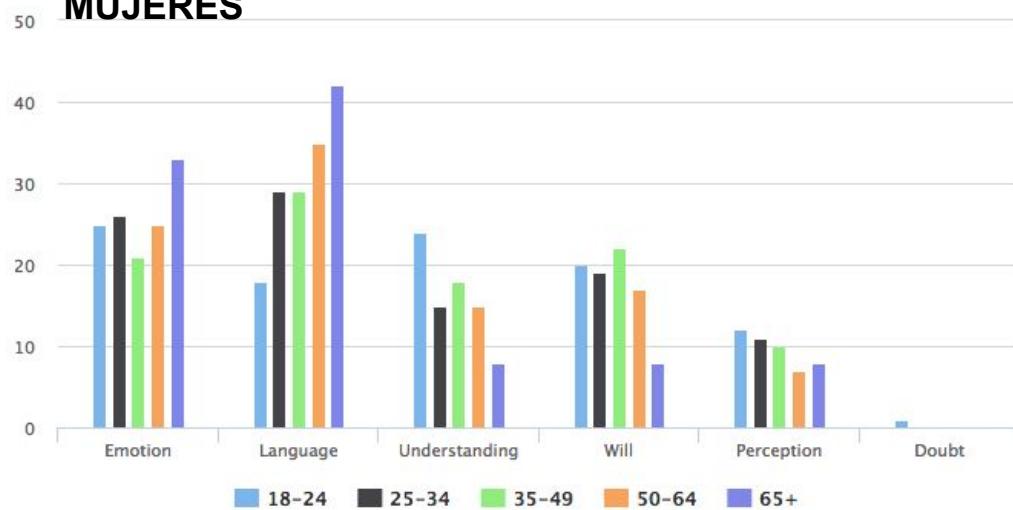
Tipos de verbos por sexo (PAN-AP14 BL-ES)

- Las mujeres usan más verbos de emoción (sentir, querer, amar...).
- Los hombres usan más verbos relativos al entendimiento (entender, comprender, estudiar...) y el permiso (poder, permitir, deber...).

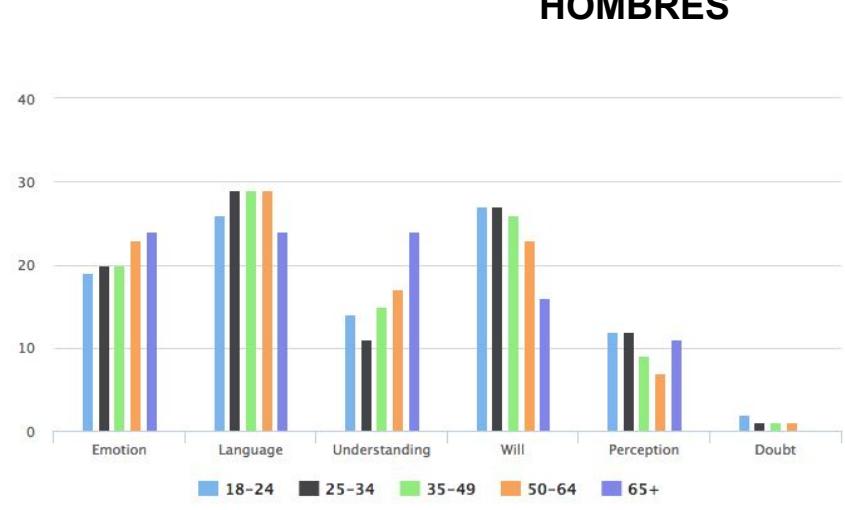


Evolución t. de verbos por sexo (PAN-AP14 BL-ES)

MUJERES



HOMBRES



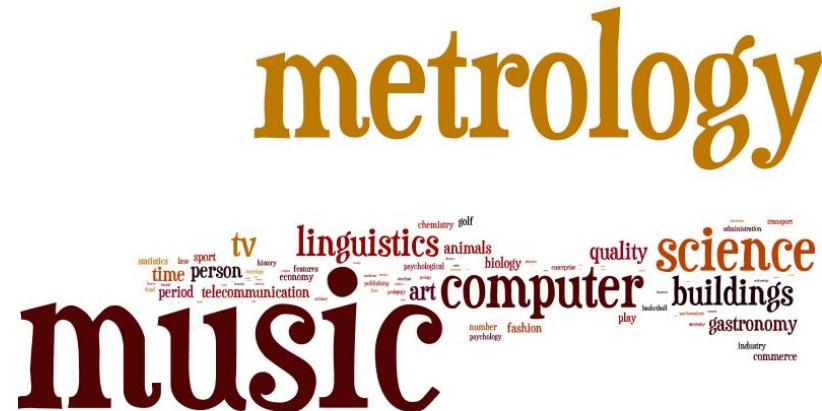
Temas por sexo (PAN-AP14 TW-EN)

MUJERES



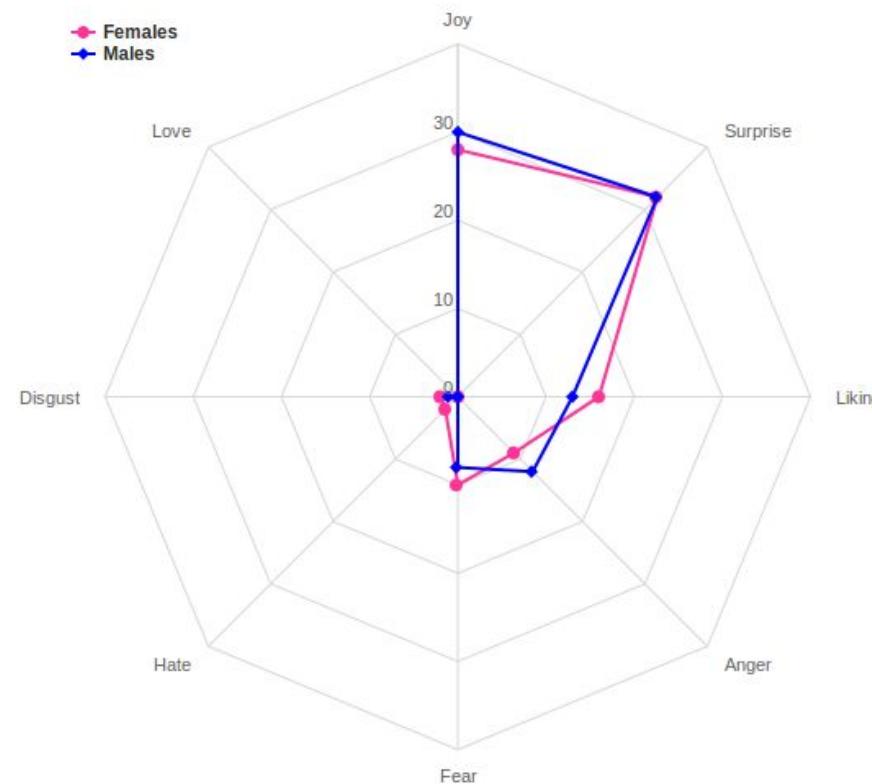
metrology

HOMBRES



- Sin diferencias significativas entre sexos.

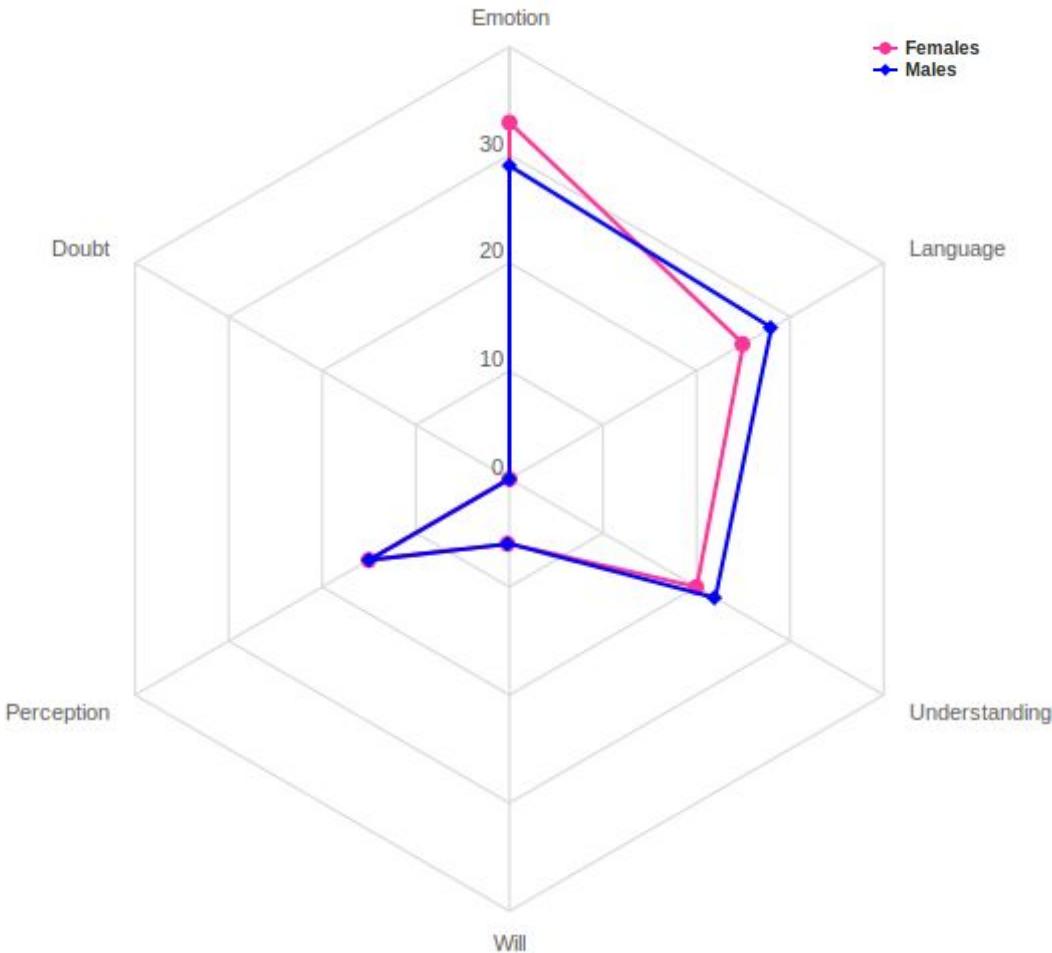
Emociones por sexo (PAN-AP14 TW-EN)



- Los hombres expresan más alegría y enfado.
- Las mujeres expresan más gusto, odio y miedo.

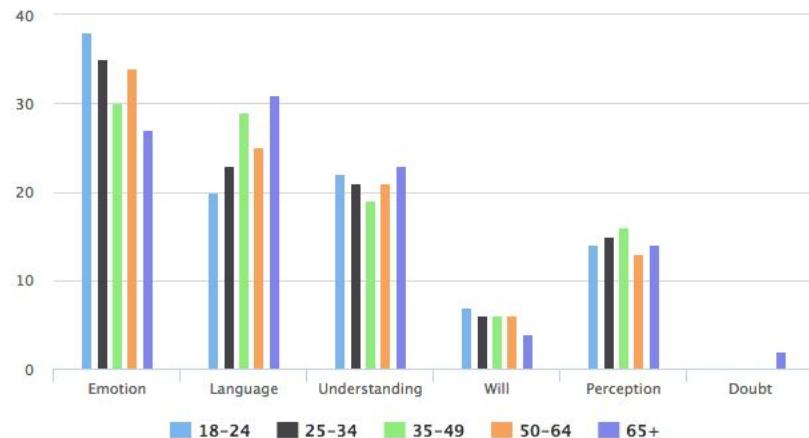
Tipos de verbos por sexo (PAN-AP14 TW-EN)

- Las mujeres usan más verbos de emoción (sentir, querer, amar...).
- Los hombres usan más verbos relativos al lenguaje (decir, contar, hablar...), y al entendimiento (entender, estudiar, comprender...).

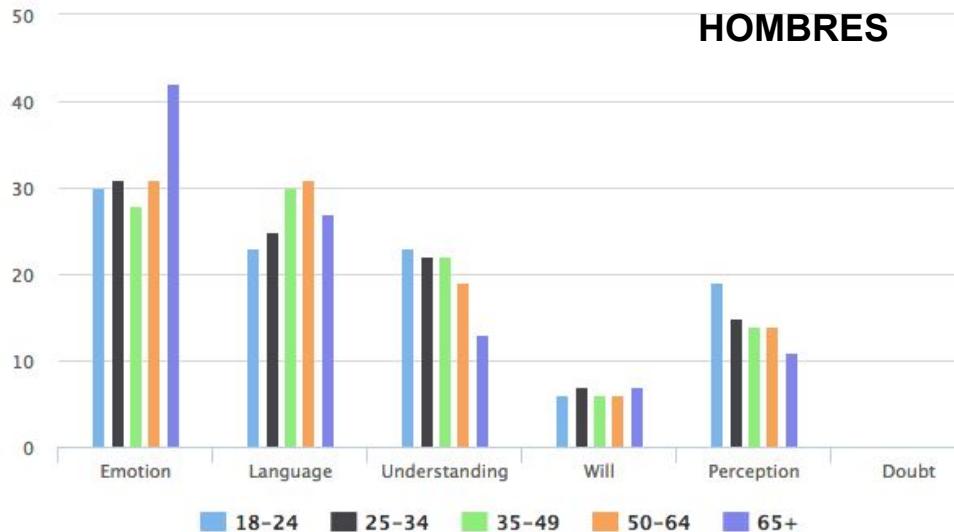


Evolución t. de verbos por sexo (PAN-AP14 TW-EN)

MUJERES



HOMBRES



Temas por sexo (PAN-AP14 TW-ES)

MUJERES

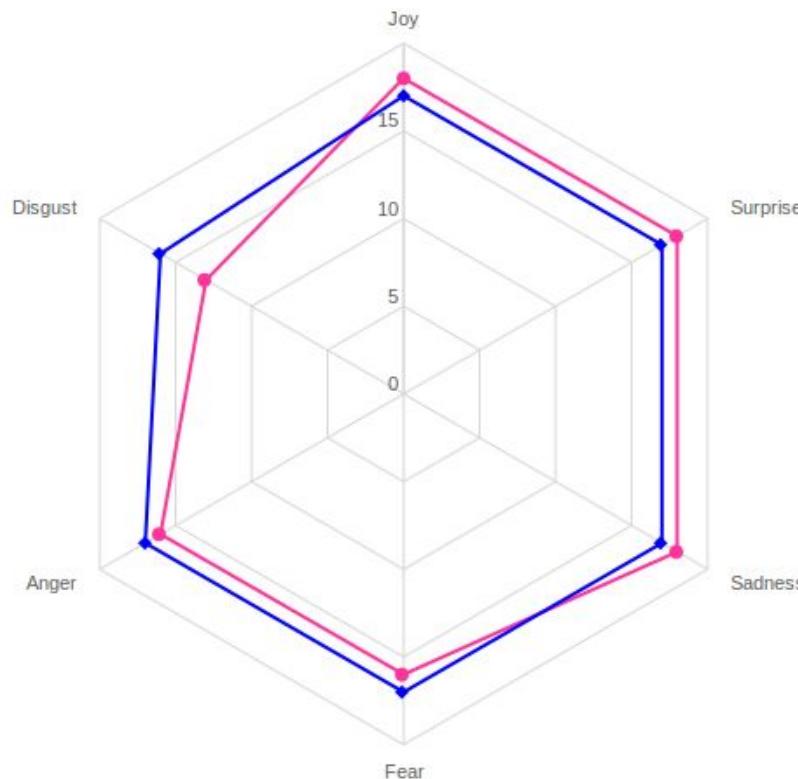


HOMBRES



- Sin diferencias significativas entre sexos.
 - Las mujeres hablan ligeramente más de la economía.

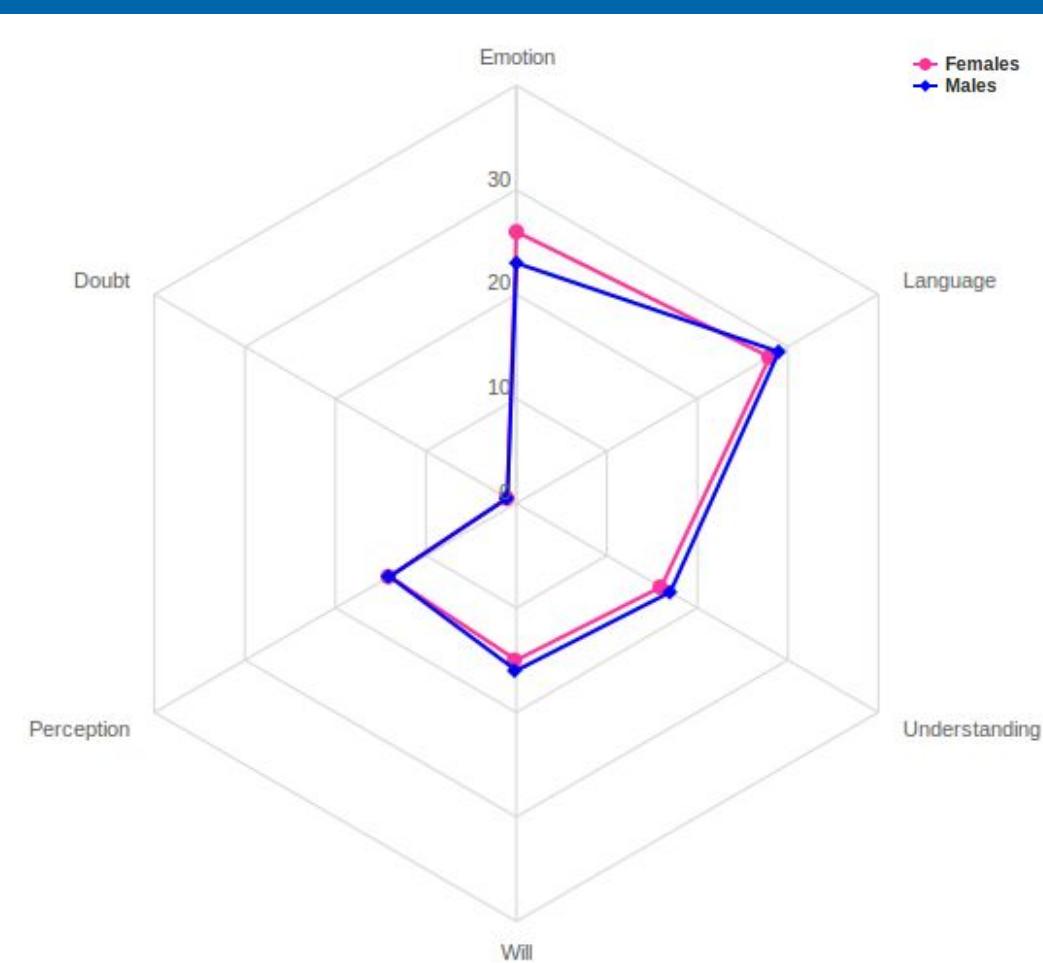
Emociones por sexo (PAN-AP14 TW-ES)



- Los hombres utilizan más disgusto, enfado y miedo, miedo.
- Las mujeres utilizan más alegría, sorpresa y tristeza.

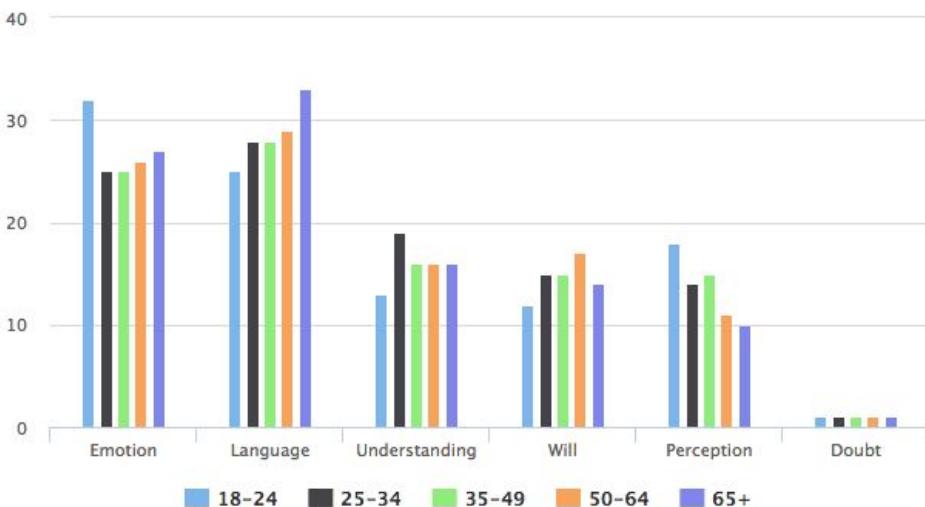
Tipos de verbos por sexo (PAN-AP14 TW-ES)

- Las mujeres usan más verbos de emoción (sentir, querer, amar...).

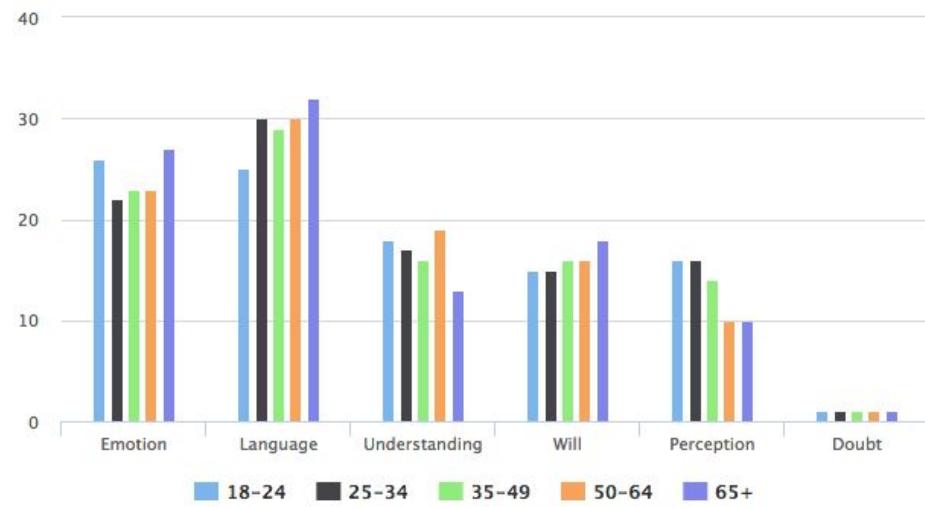


Evolución t. de verbos por sexo (PAN-AP14 TW-ES)

MUJERES



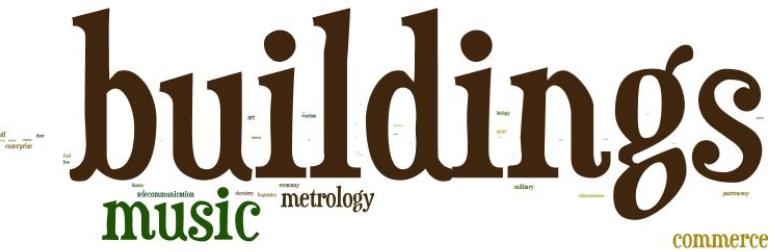
HOMBRES



Temas por sexo (PAN-AP14 RV-EN)

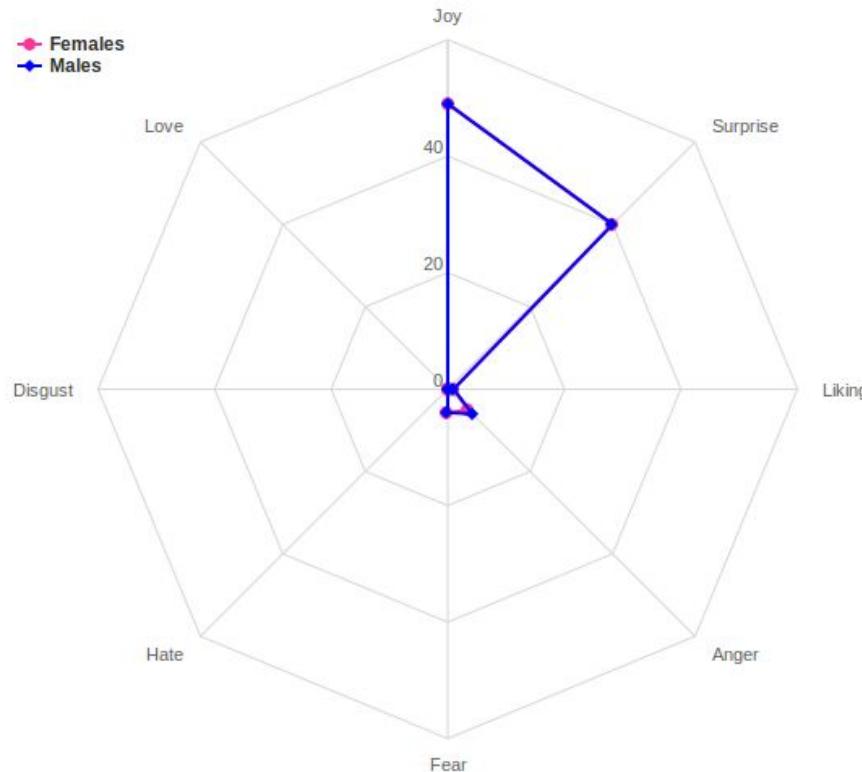
MUJERES

HOMBRES



- Sin diferencias significativas entre sexos.
- Se habla en el dominio de las construcciones (revisiones de hotel).

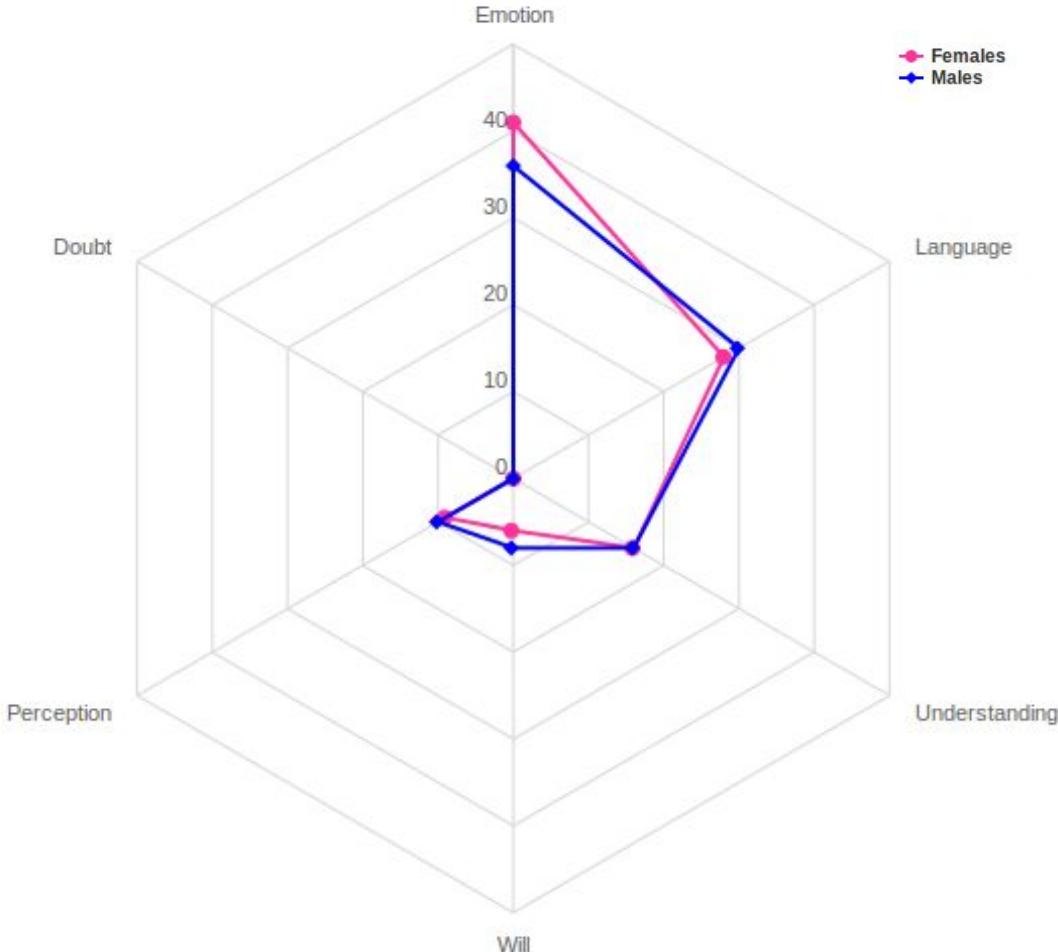
Emociones por sexo (PAN-AP14 RV-EN)



- Sin diferencias significativas entre sexos.

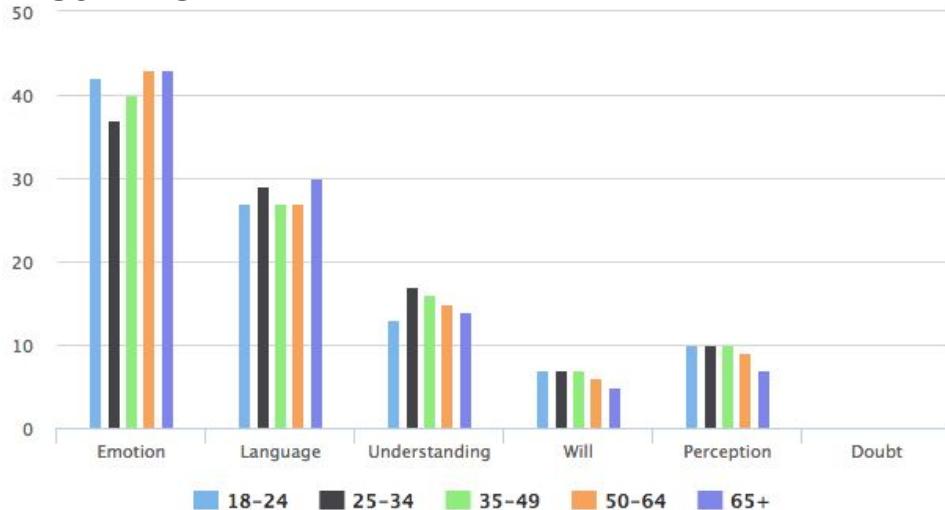
Tipos de verbos por sexo (PAN-AP14 RV-EN)

- Las mujeres usan más verbos de emoción (sentir, querer, amar...).
- Los hombres usan más verbos relativos al lenguaje (decir, contar, hablar...), o permiso (poder, deber...).



Evolución t. de verbos por sexo (PAN-AP14 RV-EN)

MUJERES



HOMBRES

