On the Multilingual and Genre Robustness of EmoGraphs for Author Profiling in Social Media

Francisco Rangel and Paolo Rosso

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

autoritas

Author profiling aims at identifying different traits such as age and gender of an author on the basis of her writings. We propose the novel EmoGraph graph-based approach where morphosyntactic categories are enriched with semantic and affective information following the next steps.

1 Morpho-syntactic annotation with Freeling

He	estado	tomando	cursos	en_línea
VAIP1S0	VAP00SM	VMG0000	NCMP000	RG

sobre	temas	valiosos	que	disfruto	estudiando
SPS00	NCMP000	AQ0MP0	PR0CN000	VMIP1S0	VMG0000

y	que	podrían	ayudarme	a	hablar	en
CC	PR0CN000	VMIC3P0	VMN0000	SPS00	VMN0000	SPS00

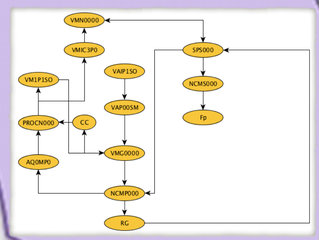
público	.
NCMS000	Fp

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

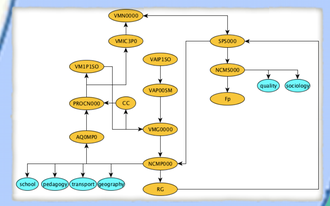
(EN) I have been taking online courses about valuable subjects that I enjoy studying and might help me to speak in public.

2 Graph creation POS sequence

VAIP1S0 → VAP00SM → VMG0000 → NCMP000 → RG → SPS00 → ...



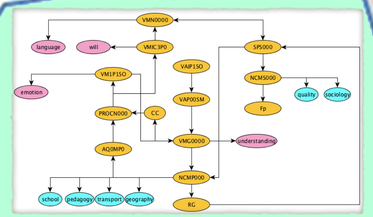
3 Topics with WordNet Domains



cursos NCMP000 transport geography pedagogy school

público NCMS000 sociology quality

4 Verbs



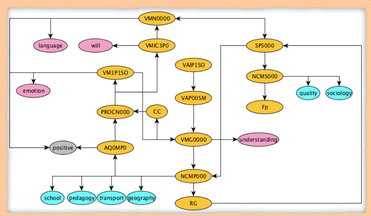
podrían VMIC3P0 will

hablar VMN0000 language

estudiando VMG0000 understanding

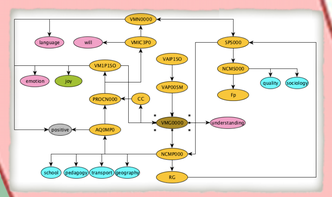
disfruto VMIP1S0 emotion

5 Polarity words



valiosos disfruto ayudarme AQ0MP0 VMIP1S0 VMN0000 positive

6 Emotions with Spanish Emotion Lexicon



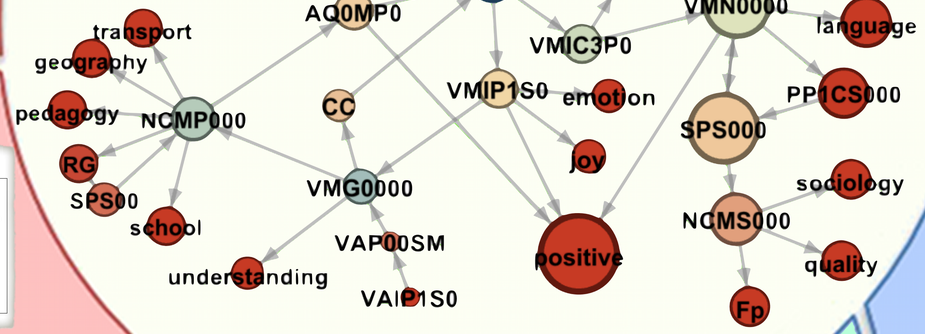
disfruto VMIP1S0 joy

7 Features

EigenVector	Given a graph and its adjacency matrix $A = a_{n,t}$ where $a_{n,t}$ is 1 if a node n is linked to a node t, and 0 otherwise: $x_n = \frac{1}{\lambda}\sum_{t\in M(n)} x_t = \frac{1}{\lambda}\sum_{t\in G} a_{n,t}x_t$ where λ is a constant representing the greatest eigenvalue associated with the centrality measure.
Betweenness	It is the ratio of all shortest paths from one node to another node in the graph that pass through x: $BC(x) = \sum_{i,j\in N-\{n\}} \frac{\sigma_{i,j}(n)}{\sigma_{i,j}}$ Where $\sigma_{i,j}$ is the total number of shortest paths from node i to j, and $\sigma_{i,j}(n)$ is the total number of those paths that pass through n.

Nodes-edges ratio	$max(E) = N * (N-1)$
Average degree Weighted average degree	Averaging all nodes degrees. Scaling it to [0,1]
Diameter	$d = max_{v\in N} e(N)$ where E(N) is the eccentricity
Density	$D = \frac{2*\|E\|}{(\|N\|*(\|N\|-1))}$
Modularity	Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E. Fast unfolding of communities in large networks. In: Journal of Statistical Mechanics: Theory and Experiment, vol. 2008 (10), pp. 10008 (2008)
Clustering coefficient	Watts-Strogatz: $cc1 = \frac{\sum_{v\in G} C(v)}{n}$
Average path length	Brandes, U. A Faster Algorithm for Betweenness Centrality. In: Journal of Mathematical So- ciology 25(2), pp. 163-177 (2001)

+ char 6-gram



Resources

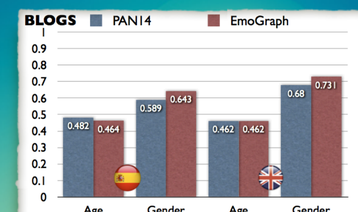
Freeling	http://nlp.lsi.upc.edu/freeling/
WordNet Domains (+EuroWordnet)	http://wndomains.fbk.eu/ http://www.illc.uva.nl/EuroWordNet/
Semantic Classification of Verbs	Levin, B. English Verb Classes and Alternations. University of Chicago Press, Chicago. (1993) a) perception (see, listen, smell...); b) understanding (know, understand, think...); c) doubt (doubt, ignore...); d) language (tell, say, declare, speak...); e) emotion (feel, want, love...); f) and will (must, forbid, allow...)
Polarity Lexicon	Hu, M., Liu, B. Mining and Summarizing Customer Reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Seattle, Wash- ington, USA, pp. 168-177 (2004)
Spanish Emotion Lexicon	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 Mex- ican International Conference on Artificial Intelligence, MICAI, pp. 1-14 (2012)

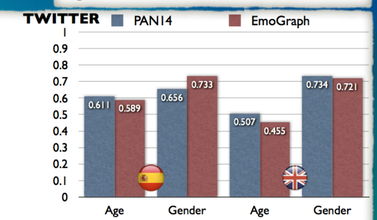
CLEF 2015 Toulouse

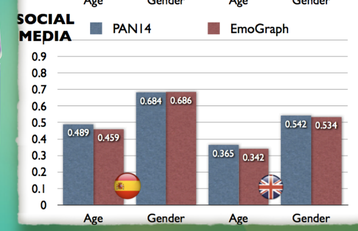
PAN-AP14 dataset

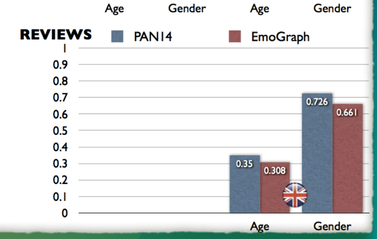
	Social Media		Blogs		Twitter		Reviews
	English	Spanish	English	Spanish	English	Spanish	English
18-24	680	150	10	4	12	4	74
25-34	900	180	24	12	56	26	200
35-49	980	138	32	26	58	46	200
50-64	790	70	10	10	26	12	200
65+	26	28	2	2	2	2	147
Σ	3376	566	78	54	154	90	821

Age and Gender Identification

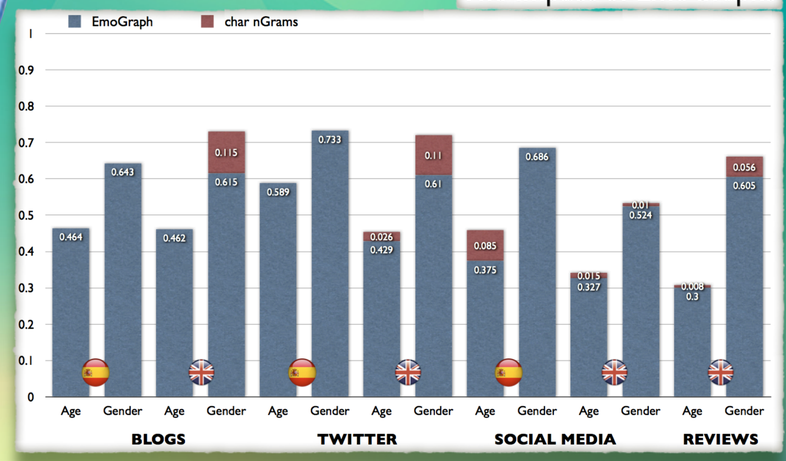








The Impact of EmoGraphs



Conclusions

EmoGraph remains robust against different:

*Genres: Blogs, Twitter, Social Media and Reviews
*Languages: Spanish, English

More competitive in Spanish

Future Work

As future work we plan to investigate further the application of:

cost-sensitive machine learning techniques