El uso del lenguaje en los diferentes canales de Internet

Francisco Manuel Rangel Pardo
Autoritas / UPV
http://www.kicorangel.com

Paolo Rosso
UPV
http://www.dsic.upv.es/~prosso/

autoritas
nuevas ideas, nuevas soluciones

VNIVERSITAT POLITÈCNICA VALÈNCIA

comunica2.0
III Congreso universitario sobre redes sociales Campus Gandia

fundéu BBVA

INTRODUCCIÓN

- El ***estilo discursivo*** es un reflejo de la ***personalidad*** del sujeto que lo elabora. La elección de las palabras y el modo en que se combinan, aporta información de dimensiones como el ***género***, la ***edad*** e incluso el ***estado emocional*** de quién las emite. Pero en Comunicación 2.0 nos enfrentamos a gran variedad de canales y surge la pregunta, ***¿define el canal el modo en que se usa el lenguaje?***

MARCO TEÓRICO

- ✓ **The Secret Life of Pronouns**. James W. Pennebaker
  - ✓ Palabras de contenido 99,96% vs Palabras de función 0,04%
  - ✓ Palabras de función
    - ✓ Cortas y difíciles de detectar
    - ✓ Muy frecuentes
    - ✓ Muy, muy sociales
    - ✓ El cerebro las procesa de manera diferente a las palabras de contenido
- ✓ **Frecuencias del Español. Diccionario y estudios léxicos y morfológicos.** Almela, R., P. Cantos, A. Sánchez, R. Sarmiento, M. Almela
  - ✓ Palabras de contenido 96,92% vs Palabras de función 3.08%
  - ✓ Sustantivos: 54%; Verbos: 22%; Adjetivos: 18%

autoritas
nuevas ideas, nuevas soluciones

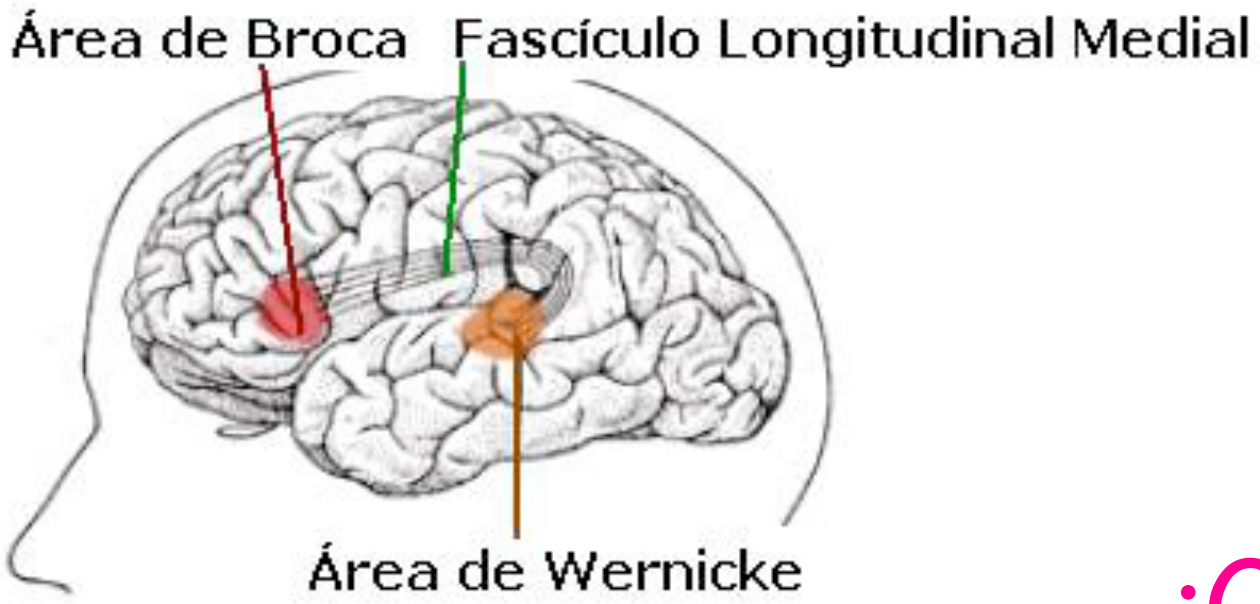
VNIVERSITAT POLITÈCNICA VALÈNCIA

comunica2.0
III Congreso universitario sobre redes sociales Campus Gandia

fundéu BBVA

MARCO TEÓRICO

¿Cómo?



¿Qué?

autoritas
nuevas ideas, nuevas soluciones

VNIVERSITAT POLITÈCNICA VALÈNCIA

comunica2.0
III Congreso universitario sobre redes sociales Campus Gandia

fundéu BBVA

METODOLOGÍA

- ✓ Se ha determinado un conjunto de ***canales*** de información por sus características representativas de diferentes colectivos de usuarios de Internet: Wikipedia, prensa, blogs, foros, twitter y facebook
- ✓ Se ha recopilado un conjunto significativo de ***documentos*** de dichos canales, todos para el idioma ***Español***
- ✓ Se ha determinado un conjunto de ***categorías gramaticales*** a analizar por su función sintáctica
- ✓ Se ha procesado mediante un POSTagger la extracción automática de la terminología en su correspondiente categoría gramatical
- ✓ Todo lo anterior se ha realizado con la herramienta ***Cosmos*** de ***Autoritas***

autoritas®
nuevas ideas, nuevas soluciones

comunica2.0
III Congreso universitario sobre redes sociales Campus Gandia

fundéu BBVA

METODOLOGÍA: CANALES

DOCUMENTOS	TÉRMINOS
3.987.179	267.465.810

DOCUMENTOS	TÉRMINOS
5.191.694	499.477.658

DOCUMENTOS	TÉRMINOS
1.083.709	122.509.753

DOCUMENTOS	TÉRMINOS
673.664	21.026.388

DOCUMENTOS	TÉRMINOS
23.873.371	163.188.448

DOCUMENTOS	TÉRMINOS
576.723	28.974.716

autoritas
nuevas ideas, nuevas soluciones

VNIVERSITAT POLITÈCNICA VALÈNCIA

comunica2.0
III Congreso universitario sobre redes sociales Campus Gandia

fundéu BBVA

METODOLOGÍA: CATEGORÍAS GRAMATICALES

- ✓ Adjetivo
- ✓ Adverbio
- ✓ Conjunción
- ✓ Cuantitativo
- ✓ Determinante
- ✓ Interjección
- ✓ Marcador discursivo
- ✓ Preposición
- ✓ Pronombre
- ✓ Sustantivo
- ✓ Verbo

✓ Se toma la persona y el número de verbos y pronombres

RESULTADOS: TÉRMINOS ÚNICOS

ÚNICOS	RATIO
162.357	1,89

ÚNICOS	RATIO
157.457	1,83

ÚNICOS	RATIO
162.412	1,89

ÚNICOS	RATIO
93.145	1,08

ÚNICOS	RATIO
128.147	1,49

ÚNICOS	RATIO
110.040	1,28

✓ Lexicon de la RAE: 85.918

autoritas
nuevas ideas, nuevas soluciones

VNIVERSITAT POLITÈCNICA VALÈNCIA

comunica2.0
III Congreso universitario sobre redes sociales Campus Gandia

fundéu BBVA

RESULTADOS: MEDIA TÉRMINOS POR DOCUMENTO

MEDIA
67

MEDIA
96

MEDIA
113

MEDIA
31

MEDIA
7

MEDIA
50

autoritas
nuevas ideas, nuevas soluciones

VNIVERSITAT POLITÈCNICA VALÈNCIA

comunica2.0
III Congreso universitario sobre redes sociales Campus Gandia

fundéu BBVA

RESULTADOS: CATEGORÍAS GRAMATICALES

CAT	WIKI	PRENSA	BLOG	FORO	TW	FB
ADJ	13,57%	12,50%	13,67%	**9,27%**	**6,62%**	12,06%
ADV	2,78%	3,46%	3,87%	4,74%	**6,30%**	3,49%
CONJ	1,52%	2,10%	1,80%	4,18%	**7,00%**	2,64%
Q	3,34%	4,47%	4,15%	5,34%	5,53%	4,29%
DET	2,88%	3,48%	2,78%	4,18%	**6,40%**	4,02%
INTJ	0,35%	0,04%	0,06%	0,42%	0,38%	0,07%
MD	0,01%	0,03%	0,02%	0,00%	0,00%	0,00%
PREP	4,00%	5,49%	5,07%	**8,94%**	**13,81%**	6,15%
PRON	0,65%	0,92%	1,12%	2,22%	**3,32%**	1,39%
NOM	**50,33%**	47,05%	46,59%	42,63%	**34,08%**	47,04%
VERB	20,55%	20,47%	20,88%	18,08%	**16,56%**	18,83%

autoritas
nuevas ideas, nuevas soluciones

comunica2.0
III Congreso universitario sobre redes sociales Campus Gandia

fundéu BBVA

RESULTADOS: PRONOMBRES Y VERBOS

CAT	PER	NUM	WIKI	PRENSA	BLOG	FORO	TW	FB
PRON	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	4,58%	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%
VERB	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%

autoritas
nuevas ideas, nuevas soluciones

VNIVERSITAT POLITÈCNICA VALÈNCIA

comunica2.0
III Congreso universitario sobre redes sociales Campus Gandia

fundéu BBVA

RESULTADOS: PALABRAS MÁS FRECUENTES

WIKI	PRENSA	BLOG	FORO	TW	FB
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**

autoritas
nuevas ideas, nuevas soluciones

comunica2.0
III Congreso universitario sobre redes sociales Campus Gandia

fundéu BBVA

CONCLUSIONES Y TRABAJO FUTURO

- ✓ El presente estudio muestra la variación en el uso del lenguaje según el canal de Internet dónde se comunica
- ✓ El propio canal acentúa el uso de determinadas categorías gramaticales que a su vez son identificativas de claves de personalidad, como el uso de la primera persona, o el uso de las preposiciones
- ✓ Las tablas de distribución deben permitir incorporar un agente correctivo en los estudios de personalidad basados en textos
- ✓ La investigación continúa mediante la división de las categorías en un nivel de detalle mayor, por ejemplo, verbos transitivos, intransitivos, copulativos, auxiliares...
- ✓ El trabajo presentado forma parte de un conjunto de trabajos en identificación de edad y género, extracción de emociones y perfiles de usuario

TRABAJOS RELACIONADOS

- ✓ Competición Author Profiling en PAN2013 (CLEF conference)

  Dado un documento, la tarea consiste en detectar el género y la edad del autor

  http://pan.webis.de

- ✓ Línea de investigación doctoral: Análisis de Emociones y Perfiles de Autor

  Seguir en http://www.kicorangel.com ó @kicorangel

- ✓ Proyecto Cosmos parcialmente financiado por:

ITC/464/2008, TSI-020100-2011-156 y IPT-2012-1220- 430000

autoritas
nuevas ideas, nuevas soluciones

comunica2.0
III Congreso universitario sobre redes sociales Campus Gandia

fundéu BBVA