
Kico’s stairway to PhD

Work #1: Language use in the different Internet channels

Work #2: Automatic Identification of Emotions in Facebook comments

Work #1: 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

comunica2.0
III Congreso universitario sobre redes sociales Campus Gandia

fundéu BBVA

INTRODUCCIÓN

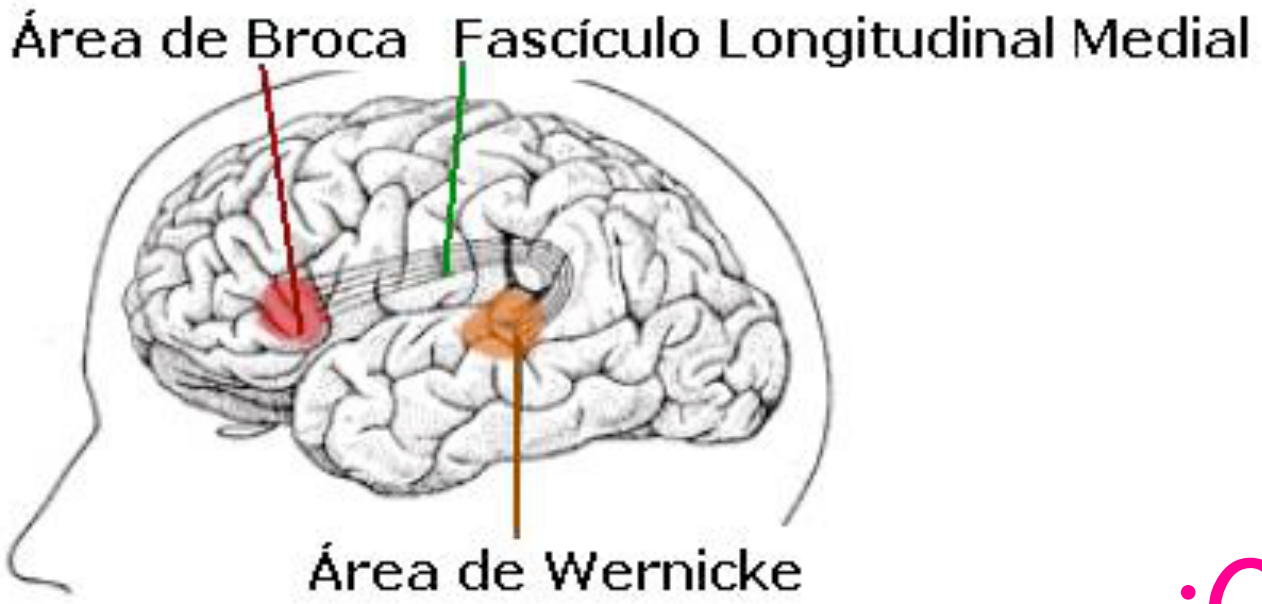
- El ***estilo discursivo*** es un reflejo de la ***personalidad*** del sujeto que lo elabora. La elección de las palabras y el modo en que se combinan, aporta información de dimensiones como el ***género***, la ***edad*** e incluso el ***estado emocional*** de quién las emite. Pero en Comunicación 2.0 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%

MARCO TEÓRICO

¿Cómo?



¿Qué?

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***

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

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

RESULTADOS: MEDIA TÉRMINOS POR DOCUMENTO

MEDIA
67

MEDIA
96

MEDIA
113

MEDIA
31

MEDIA
7

MEDIA
50

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%

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%

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**

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

Work #2: Identificación de Emociones en Facebook

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

ESTADO DEL ARTE

- ✓ Trabajos centrados en el inglés
- ✓ Estudios realizados sobre textos tradicionales: prensa, cuentos...
- ✓ Aproximaciones basadas en características semánticas muy ligadas al contexto o tema

COLECCIÓN: RECUPERACIÓN

- ✓ Recuperación de comentarios de tres temáticas de actualidad y gran activismo: política (PSOE, PP, IU, UPyD), fútbol (Real Madrid, Barça, Valencia, Athletic de Bilbao) y personajes públicos (Belén Esteban, Kiko Hernández, David Bisbal y Santiago Segura)

COLECCIÓN: RECUPERACIÓN: DOCUMENTOS

TEMA	PÁGINAS	POSTS	COMENTARIOS
POLÍTICA	PSOE	1.000	22.096
	PP	1.000	4.590
	IU	1.002	2.867
	UPyD ESTUDIANTES	593	135
FÚTBOL	REAL MADRID	1.125	1.035
	BARCELONA FC	1.002	1.520
	VALENCIA FB	1.003	463
	ATHLETIC BILBAO	560	444
PERSONAJES	BELÉN ESTEBAN	1.000	12.191
	SANTIAGO SEGURA	1.000	99
	DAVID BISBAL	1.000	2.902
	KIKO HERNÁNDEZ	1.007	291

COLECCIÓN: ETIQUETADO GÉNERO

- ✓ Con el fin de poder enlazar el estudio de las emociones con el género de los autores, se ha obtenido el género para todos los autores que lo tienen público en su muro, y se procederá a seleccionar la muestra de trabajo a partir de aquellos que lo tengan informado, balanceando entre hombres y mujeres

COLECCIÓN: SELECCIÓN DE MUESTRA

- ✓ Se balancea el mismo número de documentos para cada temática y dentro de cada temática para cada género
- ✓ No se procesan los documentos para no sesgar su selección, sólo se obtienen aquellos que contienen texto en español, aunque se realiza de manera automática y es posible que se cuelen comentarios en algún otro idioma o sólo con urls

TEMA	GÉNERO	COMENTARIOS
POLÍTICA	HOMBRES	200
	MUJERES	200
FÚTBOL	HOMBRES	200
	MUJERES	200
PERSONAJES	HOMBRES	200
	MUJERES	200
TOTAL		1.200

ETIQUETADO: FORMATO

- ✓ Se abre una hoja de excel por cada etiquetador
- ✓ En la primera celda se inserta el comentario y se abre una celda por cada una de las emociones básicas
- ✓ El etiquetador puede seleccionar tantas emociones como considere oportunas
- ✓ Si no considera ninguna emoción, deja en blanco la elección
- ✓ Ejemplo:

comentario	alegría	enfado	miedo	repulsión	sorpresa	tristeza
no tienen vergüenza. Roban y roban a manos llenas y aun encima, se hacen las víctimas.		X		X		

ETIQUETADO: REGLAS

- ✓ Tres formas de entender las emociones en textos:
  - ✓ La emoción con que debería leerse
  - ✓ Qué hace sentir
  - ✓ ***La emoción que se describe o manifiesta en la oración***
- ✓ Para evitar la implicación emocional del etiquetador y la subjetividad dentro de lo posible, se debe realizar el etiquetado intentando seguir el tercer modo de entender las reglas.

ETIQUETADO: REGLAS

- ✓ A continuación las emociones básicas y las secundarias más cercanas para ayudar al etiquetado:

ALEGRÍA	ENFADO	MIEDO	REPULSIÓN	SORPRESA	TRISTEZA
Agradecido	Agresivo	Acomplejado	Aborrecimiento	Extrañeza	Abatido
Alegre	Colérico	Alarmado	Desagrado	Sobresalto	Agobiado
Animado	Crispado	Angustiado	Grima	Susto	Apenado
Calmado	Descontento	Ansioso	Repulsión	Consternación	Confuso
Confiado	Enfadado	Atemorizado	Antipatía	Pasmo	Decepcionado
Contento	Enojado	Aterrado	Aversión	Desconcierto	Deprimido
Dichoso	Excitado	Avergonzado	Repugnancia	Estupor	Desalentado
Encantado	Fastidiado	Confuso	Disgusto	Asombro	Desanimado
Entusiasmado	Furioso	Desesperado	Repudia	Fascinación	Desdichado
Eufórica	Insatisfecho	Desorientado	Repulsa	Admiración	Desmoralizado
Esperanzado	irascible	Horrorizado	Odio	Confusión	Frustrado
Feliz	Malhumorado	Inquieto	Manía	Chasco	Nostálgico
Gozoso	Molesto	inseguro	Rabia	Impresión	Soledad
Satisfecho	Nervioso	Intranquilo	Animadversión	Exclamación	Triste
Tranquilo	Rabioso	Pánico	Nauseabundo	Conmoción	Infeliz
Complacido	Tenso	Preocupado	Indignación	Estupefacción	Desconsolado
Libre	Violento	Temeroso	Enfado		Afligido
Fascinado	Irritado	Tenso	Desprecio		Amargado
Seguro	Indignado	Indeciso	Distanciamiento		Impotente
		Impotencia			

ETIQUETADO: EVALUACIÓN

- ✓ Dependiendo del número de etiquetadores y de la cantidad de emociones que se pueden etiquetar por texto, tenemos:
  - ✓ Kappa de Kohen: Dos etiquetadores, una emoción por texto
  - ✓ Kappa de Fleiss: Múltiples etiquetadores, una emoción por texto
  - ✓ Kappa de Kohen ponderada: Dos etiquetadores, múltiples emociones por texto
  - ✓ **Kappa de Díaz-Sidorov: Múltiples etiquetadores, múltiples emociones por texto**

CONCORDANCIA

BÁSICAS	KICO	MERCHE	ROSA	RESTO
KICO		0,0587	0,2738	0,1662
MERCHE	0,0587		0,1042	0,0814
ROSA	0,2738	0,1042		0,1890
TOTAL	0,1455			

COMBINADAS	KICO	MERCHE	ROSA	RESTO
KICO		0,6618	0,5656	0,6137
MERCHE	0,6618		0,5773	0,6196
ROSA	0,5656	0,5773		0,5715
TOTAL	0,6016			

ETIQUETADO: SELECCIÓN

- ¿Qué emoción se selecciona para cada comentario?
  - La que haya sido marcada por cualquier anotador (1/3)
  - **La que haya sido marcada por la mayoría de anotadores (2/3)**
  - La que haya sido marcada por todos los anotadores (3/3)

RESULTADOS DEL ETIQUETADO (2/3)

BÁSICAS	ALEGRÍA	ENFADO	MIEDO	REPULSIÓN	SORPRESA	TRISTEZA	NEUTRA
TOTAL	338	151	3	129	390	76	262
%	28,17	12,58	0,25	10,75	32,50	6,33	21,83

COMBINADAS	ALEGRÍA / SORPRESA	ENFADO / REPULSIÓN	MIEDO	TRISTEZA	NEUTRA
TOTAL	639	243	3	76	262
%	53,25	20,25	0,25	6,33	21,83

MODELO APRENDIZAJE: CARACTERÍSTICAS

(F)recuencias: Ratio frente al total del número de palabras únicas, palabras que empiezan en mayúsculas, palabras completamente en mayúsculas, longitud de las palabras, caracteres en mayúsculas y palabras alargadas (pe. Holaaa).
(P)untuación: Frecuencia de uso de puntos, comas, puntos y comas, dos puntos, exclamaciones, interrogaciones y número de comillas.
(C)ategorías gramaticales utilizadas (PoS Tagger). Número y persona de verbos y pronombres, modos verbales, nombres propios (NER) y palabras no identificadas.
(E)moticonos, que aún definiendo el estilo de escritura, incorpora semántica de contenido emocional (Martínez et al., 2012) pero manteniendo independencia del contexto temático (Read, 2005). Ratio del número de emoticonos frente al total de palabras, número de emoticonos alegres, enfadados, disgustados, sorprendidos, triste, burla y mudos[6]

(SEL) Spanish Emotion Lexicon (Sidorov et al., 2012): Para cada palabra del comentario se obtiene su lema y su FPA en el diccionario. Si el lema no tiene entrada, se obtienen sus sinónimos y se obtiene el FPA de los mismos. Se suman todos los FPA para cada emoción.
(BoW): Se obtiene el lema de las 20 primeras palabras con mayor ganancia de información y tratando que sean lo más independientes de la temática posible, para lo que se han obtenido principalmente adjetivos y adverbios, eliminando de estas palabras los sustantivos (pe. gol, recorte...). Se han reducido a su raíz las palabras alargadas (pe. Holaaa) y se ha reducido cualquier combinación de risas (ja, je, ji, jo, ju) y todos sus alargamientos (jajaaj, jejejeje) a una característica común: ja.

ENTRENAMIENTO Y EVALUACIÓN

- ✓ Cuatro algoritmos de aprendizaje
  - ✓ Support Vector Machines
  - ✓ Decission trees
  - ✓ Naïve Bayes
  - ✓ Bayes Net
- ✓ Dos evaluaciones como en SEMEVal 2007
  - ✓ Precision, Recall and F
  - ✓ *r* Kappa de Pearson para correlación

EXPERIMENTOS

- ✓ Experimento 1: Uso de diccionario afectivo
- ✓ Experimento 2: Identificación basada en estilo
- ✓ Experimento 3: Emociones básicas vs. combinadas
- ✓ Experimento 4: Identificación de género

#1: Uso de SEL

		rr	Prec	Rec	F1
Ale.	FPEC	21,2	39,2	**64,5**	48,8
	+SEL	**23,6**	**70,7**	62,9	**64,8**
Enf.	FPEC	**22.98**	**84,8**	72,9	77,0
	+SEL	22,36	84,4	**73,3**	**77,2**
Rep.	FPEC	23,18	**87,1**	74,3	78,7
	+SEL	**23,88**	87,0	**75,6**	**79,6**
Sor.	FPEC	17,8	42,3	56,7	48,4
	+SEL	**18,0**	**64,7**	**60,2**	**61,4**
Tri.	FPEC	15,9	90,9	80,4	84,7
	+SEL	**16,0**	90,9	**80,5**	84,7

- ✓ El uso de diccionario mejora significativamente la identificación en la mayoría de los casos

#2: Identificación basada en estilo

		rr	Prec	Rec	F1
Ale.	J48	27,1	70,6	71,7	71,0
	NB	**27,9**	71,1	68,6	69,5
	BN	25,6	**72,5**	62,6	64,4
	SVM	24,9	71,3	**73,9**	**71,1**
Enf.	J48	16,6	81,9	**84,7**	**83,1**
	NB	**22,6**	**84,5**	73,3	77,2
	BN	22,2	**84,5**	72,8	76,8
	SVM	10,8	80,6	83,8	82,0
Rep.	J48	21,7	85,2	**87,3**	**86,1**
	NB	15,7	85,3	70,8	75,9
	BN	**24,9**	**87,3**	75,9	79,9
	SVM	6,2	81,0	85,4	83,0
Sor.	J48	**25,8**	**67,5**	**67,8**	**67,6**
	NB	20,6	66,6	60,0	61,3
	BN	20,7	66,3	60,7	61,9
	SVM	17,2	64,5	67,3	64,8
Tri.	J48	12,1	89,6	90,9	90,2
	NB	6,1	89,3	75,3	81,1
	BN	**16,7**	**91,1**	80,2	84,5
	SVM	8,2	89,2	**91,6**	**90,3**

- ✓ Las características propuestas, basadas en estilo y de independencia de la temática, proporcionan unos resultados competitivos con el estado del arte (SEMEval) para medios sociales en Español

#3: Emociones básicas vs. combinadas

		π	Prec	Rec	F1
Ale. + Sor.	J48	38,8	69,5	69,6	69,5
	NB	42,1	71,3	71,1	71,1
	BN	40,1	70,5	70,4	70,2
	SVM	**44,5**	**72,9**	**72,1**	**72,1**
Enf. + Dis.	J48	26,0	76,2	**77,3**	**76,7**
	NB	**33,9**	**80,2**	73,0	75,2
	BN	33,0	79,3	73,8	75,7
	SVM	18,9	74,2	**77,3**	75,3

- ✓ Es latente la dificultad de identificación unívoca de emociones cercanas

#4: Identificación de género

	Acc	*r*	Prec	Rec	F1
Hombre	53,6	7,67	54,2	49,7	51,8
Mujer			53,5	53,8	53,8

- ✓ El método propuesto sirve para la identificación del género del autor más allá del puro azar (r=7,67), aunque habrá que esperar para compararlo con resultados como los del PAN

Conclusiones y trabajo futuro

- ✓ Hemos generado un corpus de un medio social como son los comentarios de FAcebook y en español, etiquetado en las seis emociones de Eckman
- ✓ Hemos propuesto un método de identificación de emociones basado en características de estilo e independientes de la temática
- ✓ Hemos comprobado la dificultad de identificación de emociones cercanas, como disgusto/enfado y alegría/sorpresa

- ✓ En futuros trabajos verificaremos la propuesta con los datasets de estudios similares (Díaz) y con la competición de Author Profiling de PAN 2013
- ✓ Vamos a indagar más en profundidad en características de estilo de escritura

That's all folks! ;-)