Use of Language and Author Profiling: Identification of Gender and Age

Francisco Rangel & Paolo Rosso

Universitat Politècnica de València

autoritas
nuevas ideas, nuevas soluciones

NLPCS 2013

10th International Workshop on
Natural Language Processing and Cognitive Science

CIRM, Marseille, France - 16 October 2013

What's Author Profiling?

Gender?

Personality traits?

Age?

Emotions?

Native language?

Author Profile... Who is who?

Why Author Profiling?

Forensics	Security	Marketing
Language as evidence	***Profiling possible delinquents***	***Segmenting users***

Research Goals

- Based on our preliminary research on Spanish language we investigated further...
  - ...how the language is used in different channels of Internet (Wikipedia, newsletters, blogs, forums, Twitter, Facebook)[1]
  - ...how the language could provide enough evidences to identify the six basic emotions of Eckman[2]: joy, anger, sadness, surprise, fear, disgust

WE AIM AT MODELING THE DIFFERENCES IN THE USE OF LANGUAGE BY AGE AND GENDER

[1. El Uso del Lenguaje en los Diferentes Canales de Internet. Rangel, F., Rosso, P. In: Proc. Comunica 2.0 Gandía, Spain. Feb 21-22]

[2. On the Identification of Emotions and Authors' Gender in Facebook Comments on the Basis of their Writing Style. Rangel, F., Rosso, P. (To appear)]

Outline

- Brief review to state-of-the-art in Author Profiling
  - Special focus on Age and Gender identification
  - Author Profiling at PAN-CLEF-2013 as example
- Preliminary experiments on use of language
- Methodology
- Features
- Experimental results
- Conclusions and future work

Outline

- Brief review to state-of-the-art in Author Profiling
  - Specially focus on Age and Gender identification

Related Work on Computational Linguistics

AUTHOR	COLLECTION	FEATURES	RESULTS	OTHER CHARACTERISTICS
Argamon et al., 2002	British National Corpus	Part-of-speech	Gender: 80% accuracy	
Holmes & Meyerhoff, 2003	Formal texts	-	Age and gender	
Burger & Henderson, 2006	Blogs	Posts length, capital letters, punctuations. HTML features.	They only reported: "Low percentage errors"	Two age classes: [0,18[,[18,-]
Koppel et al., 2003	Blogs	Simple lexical and syntactic functions	Gender: 80% accuracy	Self-labeling
Schler et al., 2006	Blogs	Stylistic features + content words with the highest information gain	Gender: 80% accuracy Age: 75% accuracy	
Goswami et al., 2009	Blogs	Slang + sentence length	Gender: 89.18 accuracy Age: 80.32 accuracy	
Zhang & Zhang, 2010	Segments of blog	Words, punctuation, average words/sentence length, POS, word factor analysis	Gender: 72,10 accuracy	
Nguyen et al., 2011 y 2013	Blogs & Twitter	Unigrams, POS, LIWC	Correlation: 0.74 Mean absolute error: 4.1 - 6.8 years	Manual labeling Age as continuous variable
Peersman et al., 2011	Netlog	Unigrams, bigrams, trigrams and tetagrams	Gender+Age: 88.8 accuracy	Self-labeling, min 16 plus 16,18,25

Related Tasks

TASK	OBJECTIVE
PAN@CLEF 2013	AGE & GENDER IDENTIFICATION
BEA-8	NATIVE LANGUAGE IDENTIFICATION
IVWSM-2013	PERSONALITY RECOGNITION -> BIG FIVE THEORY
Kaggle	PSYCHOPATHY PREDICTION BASED ON TWITTER USAGE
	PERSONALITY PREDICTION BASED ON TWITTER STREAM
	GENDER PREDICTION FROM HANDWRITING

Outline

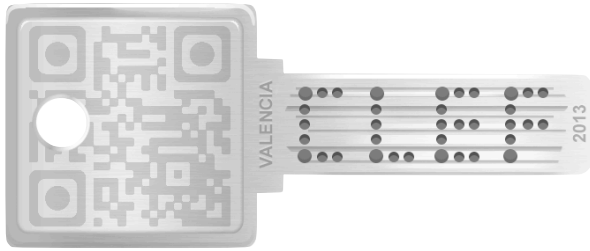
- Brief review to state-of-the-art in Author Profiling
  - Specially focus on Age and Gender identification
  - Author Profiling at PAN-CLEF-2013 as example

Author Profiling

PAN-AP-2013 - CLEF 2013

Valencia, 24th September 2013

PAN



Francisco Rangel
Autoritas / Universitat Politècnica de València

Paolo Rosso
Universitat Politècnica de València

Moshe Koppel
Bar-Illan University

Efstathios Stamatatos
University of the Aegean

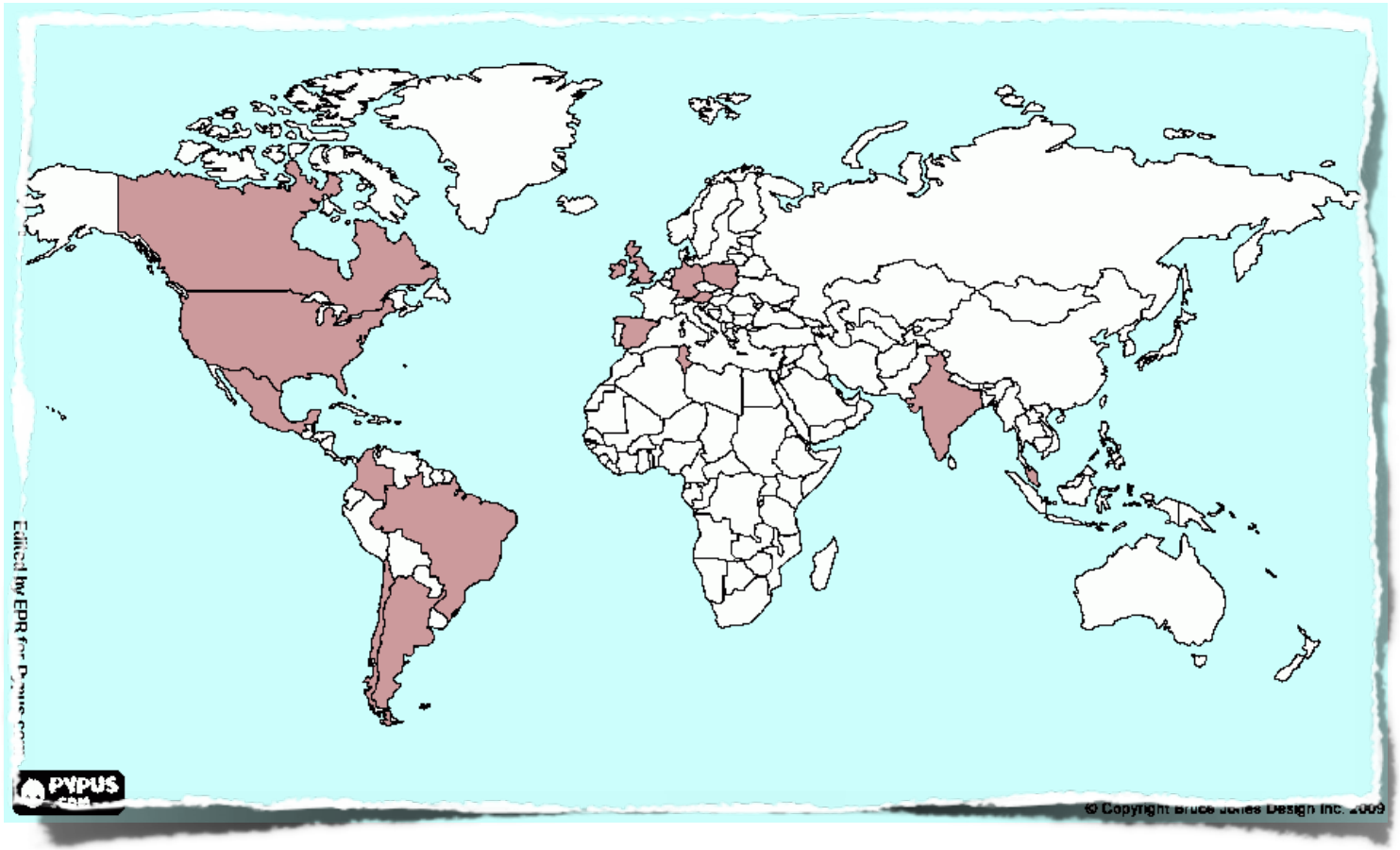
Giacomo Inches
University of Lugano

Task Main Goal

- Given a collection of documents retrieved from Social Media in English and Spanish...

MAIN GOAL
Identifying age and gender

Participants



- 66 registered teams
- 21 participants (32%)
- 16 countries
- 18 papers (86%)
  - 8 long papers
  - 10 short papers

Approaches

- What kind of ...

Preprocessing

Features

Methods

... did the teams perform?

Approaches

Preprocessing

HTML Cleaning to obtain plain text	5 teams: [gopal-patra][moreau][meina] [weren][pavan]
Deletion of documents with at least 0.1% of spam words	1 team: [flekova]
Principal Component Analysis to reduce dimensionality	1 team: [yong-lim]
Subset selection during training to reduce dimensionality	5 teams: [caurcel-diaz][flekova][moreau] [hernandez-farias][sapkota]
Discrimination between human-like posts and spam-like posts (chatbots)	1 team: [meina]

Approaches

Features

Stylistic features: frequencies of punctuation marks, capital letters, quotations...	9 teams: [yong-lim][cruz][pavan][gopal-patra][de-arteaga][meina][flekova][aleman][santosh]
+ POS tags	5 teams: [yong lim][meina][aleman][cruz][santosh]
HTML-based features like image urls or links	3 teams: [santosh][sapkota][meina]
Readability	7 teams: [gopal-patra][yong-lim][meina][flekova][aleman][weren][gillam]
Emoticons	2 teams: [aleman][hernandez-farias] *[sapkota] explicitly discarded them

Approaches

Features

Content features: LSA, BoW, TF-IDF, dictionary-based words, topic-based words, entropy-based words...	11 teams: [sapkota][gopal-patra][yong-lim][seifeddine][caurcel-diaz][flekova][meina][cruz][santosh][pavan][hernandez-farias]
Named entities	1 team: [flekova]
Sentiment words	1 team: [gopal-patra]
Emotions words	1 team: [meina]
Slang, contractions and words with character flooding	4 teams: [flekova][caurcel-diaz][aleman][hernandez-farias]

Approaches

Features

Text to be identified is used as a query for a search engine	1 team: [weren]
Unsupervised features based on statistics	1 team: [de-arteaga]
Language models (n-grams)	4 teams: [meina][jankowska][moreau] [sapkota]
Collocations	1 team: [meina]
Second order representation based on relationships between documents and profiles	1 team: [pastor]

Approaches

Methods

Decision Trees	5 teams: [santosh][gopal-patra] [seifeddine][gillam][weren]
Support Vector Machines	3 teams: [yong-lim][cruz][sapkota]
Logistic Regression	2 teams: [de-arteaga][flekova]
Naïve Bayes	1 team: [meina]
Maximum Entropy	1 team: [pavan]
Stochastic Gradient Descent	1 team: [caurcel-diaz]
Random Forest	1 team: [aleman]
Information Retrieval	1 team: [weren]

Outline

- Brief review to state-of-the-art in Author Profiling
  - Specially focus on Age and Gender identification
  - Author Profiling at PAN-CLEF-2013 as example
- Preliminary experiments on use of language

Use of Language per Channel in Spanish

- Number of documents analyzed per channel...

CHANNEL	DOCS.	TERMS	UNIQ.
Wikipedia	3,987,179	267,465,810	162,357
Newsletters	5,191,694	499,477,658	157,457
Blogs	1,083,709	122,509,753	162,412
Forums	673,664	21,026,388	93,145
Twitter	23,873,371	163,188,448	128,147
Facebook	576,723	28,974,716	110,040

Use of Language per Channel in Spanish

- Let's do a brief summary of the function of the main grammatical categories

NOUNS	Name the things
VERBS	Define the action
ADJECTIVES	Describe things, mainly complementing nouns
ADVERBS	Help describing the context, mainly complementing verbs but also other adverbs, adjectives or even the whole sentence
PREPOSITIONS	Are used to contextualize the world in a hierarchical way: Local, directional, modal, temporal, ...

Use of Language per Channel in Spanish

- Distribution of Grammatical Categories per Channel

POS	WIKI	NEWS	BLOGS	FORUMS	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	47.07
VERB	20.55	20.47	**20.88**	18.08	16.56	18.83

- TW main motto: “What’s happening?” Forum objective: Describe problem and ask for help
- Wiki, News... more descriptive channels

Use of Language per Channel in Spanish

- Frequency of Person and Number of Pronouns and Verbs

POS	PER	NUM	WIKI	NEWS	BLOG	FOR	TW	FB
PRON	1	SIN	13.61	14.58	18.85	**54.47**	**65.81**	22.3
		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
	OTHER		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

- TW and Forum are self-centered channels
- Wiki, News, ... are descriptive channels of things, people, places...

Use of Language per Channel in Spanish

- Distribution of Grammatical Categories by Gender

POS	ALL	MALE	FEMALE
ADJ	6.49	6.53	6.45
ADV	3.93	3.94	3.91
CONJ	9.51	9.55	9.46
Q	5.46	5.76	5.12
DET	7.25	6.81	**7.74**
INTJ	0.23	0.18	**0.30**
MD	0.00	0.00	0.00
PREP	6.06	**6.25**	5.85
PRON	2.45	2.24	**2.67**
NOM	31.89	32.21	31.53
VERB	15.38	15.44	15.32

+13.66%

+66.67%

+6.84%

+19.20%

- Correlates with Pennebaker's results in "The Secret Life of Pronouns"

Outline

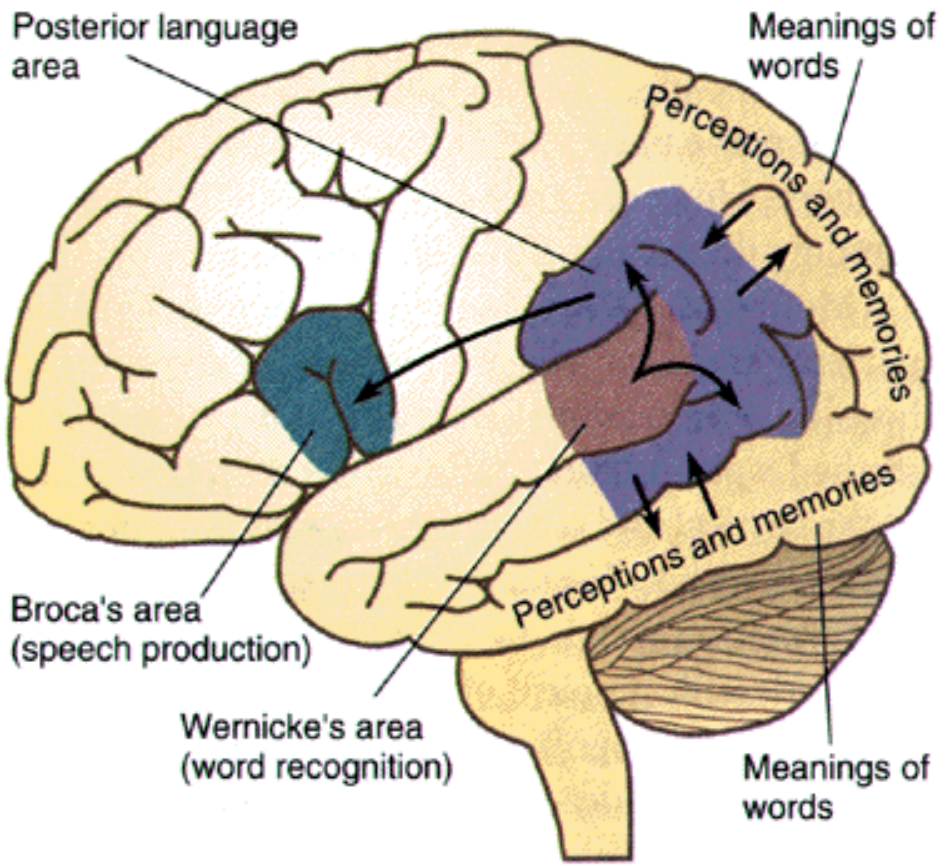
- Brief review to state-of-the-art in Author Profiling
  - Specially focus on Age and Gender identification
  - Author Profiling at PAN-CLEF-2013 as example
- Preliminary experiments on use of language
- Methodology

Theoretical Framework

- ✓ **The Secret Life of Pronouns**. James W. Pennebaker
  - ✓ Content words 99,96% vs Function words 0,04%
  - ✓ Function words
    - ✓ Short and very difficult to detect
    - ✓ High frequency
    - ✓ Very, very social
    - ✓ They are processed by the brain in a different way than content words
- ✓ **Frecuencias del Español. Diccionario y estudios léxicos y morfológicos.** Almela, R., P. Cantos, A. Sánchez, R. Sarmiento, M. Almela
  - ✓ Content words 96,92% vs Function words 3.08%
  - ✓ Nouns: 54%; Verbs: 22%; Adjectives: 18%

Neurology, a Theoretical Framework

How?



What?

Methodology

- We used the Author Profiling dataset from PAN@CLEF 2013
  - Data balanced by gender
  - Age groups: 10s (13-17), 20s (23-27), 30s (33-47)

AGE	NUM. OF AUTHORS	
	TRAIN	TEST
10s	2,500	240
20s	42,600	3,840
30s	30,800	2,720

- Machine learning approach (Weka)
  - Support Vector Machine, Gaussian kernel with g=0.01, c=2,000
- Same evaluation measure than PAN task
  - Accuracy

Outline

- Brief review to state-of-the-art in Author Profiling
  - Specially focus on Age and Gender identification
  - Author Profiling at PAN-CLEF-2013 as example
- Preliminary experiments on use of language
- Methodology
- Features

Features

PART-OF-SPEECH (GRAMMATICAL CATEGORIES)	Frequency of use of each grammatical category, number and person of verbs and pronouns, mode of verb, proper nouns (NER) and non-dictionary words (words not found in dictionary);
FREQUENCIES	Ratio between number of unique words and total number of words, words starting with capital letter, words completely in capital letters, length of the words, number of capital letters and number of words with flooded characters (e.g. Heeeelloooo);
PUNCTUATION MARKS	Frequency of use of dots, commas, colon, semicolon, exclamations, question marks and quotes;
EMOTICONS	Ratio between the number of emoticons and the total number of words, number of the different types of emoticons representing emotions: joy, sadness, disgust, angry, surprised, derision and dumb;
SPANISH EMOTION LEXICON (SEL)	We obtained the lemma for each word and then its *Probability Factor of Affective Use* value from the SEL dictionary. If the lemma does not have an entry in the dictionary, we look for its synonyms. We add all the values for each emotion, building one feature per emotion [1].

[1. On the Identification of Emotions and Authors' Gender in Facebook Comments on the Basis of their Writing Style. Rangel, F., Rosso, P. (To appear)]

Outline

- Brief review to state-of-the-art in Author Profiling
  - Specially focus on Age and Gender identification
  - Author Profiling at PAN-CLEF-2013 as example
- Preliminary experiments on use of language
- Methodology
- Features
- Experimental results

Experimental Results

- PAN ranking for Author Profiling by Gender and Age (Spanish)

POS	TEAM	GENDER	POS	TEAM	AGE
1	Santosh	0.6473	1	Pastor	0.6558
2	Pastor	0.6299	2	Santosh	0.6430
3	Haro	0.6165	**3**	**(*Rangel*)**	**0.6350**
4	Ladra	0.6138	4	Haro	0.6219
5	Flekova	0.6103	5	Flekova	0.5966
6	Jankowska	0.5846	6	Ladra	0.5727
7	**(*Rangel*)**	**0.5713**	7	Yong	0.5705
8	Kern	0,5706	8	Ramirez	0.5651
9	Jimenez	0.5627	9	Aditya	0.5643
10	Ayala	0.5526	10	Jimenez	0.5429
11	Cagnina	0.5516	11	Gillam	0.5377
12	Yong	0.5468	12	Kern	0.5375
13	Mechti	0.5455	13	Moreau	0.5049
14	Weren	0.5362	14	Meina	0.4930
15	Meina	0.5287	15	Weren	0.4615
16	Ramirez	0.5116	16	Jankowska	0.4276
17	*Baseline*	*0.5000*	*17*	Cagnina	0.4148
18	Aditya	0.5000	18	Hidalgo	0.4000
19	Hidalgo	0.5000	19	Farias	0.3554
20	Farias	0.4982	*20*	*Baseline*	*0.3333*
21	Moreau	0.4967	21	Ayala	0.2915
22	Gillam	0.4784	22	Mechti	0.0512

Outline

- Brief review to state-of-the-art in Author Profiling
  - Specially focus on Age and Gender identification
  - Author Profiling at PAN-CLEF-2013 as example
- Preliminary experiments on use of language
- Methodology
- Features
- Experimental results
- Conclusions and future work

Conclusions & Future Work

- We have analyzed a high number of documents from different channels and ...
  - ...some important variations in use of the grammatical categories by gender were appreciated
- We modeled the language only with stylistic features, independent from contents, topics, themes...
  - ... verifying that such features help to identify age and gender of anonymous authors because ...
  - ... we obtained competitive results compared to participants at PAN-AP@CLEF 2013
- As future work...
  - We plan to analyze the discourse more in depth...
    - ...for example using collocations because...
    - ...the order is very important: “She married and become pregnant vs. she become pregnant and married” Michael Zock and Debela Tesfaye
  - We want to investigate the relationship between demographics (age, gender) with the emotional and personality profiles

Thank you very much!!

Francisco Rangel
@kicorangel

Paolo Rosso
prosso@dsic.upv.es

Our main objective is to build a common framework which allows us to better understanding how people use the language and how the language helps profiling them

autoritas
nuevas ideas, nuevas soluciones

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

NLPCS 2013

10th International Workshop on Natural Language Processing and Cognitive Science

CIRM, Marseille, France - 16 October 2013