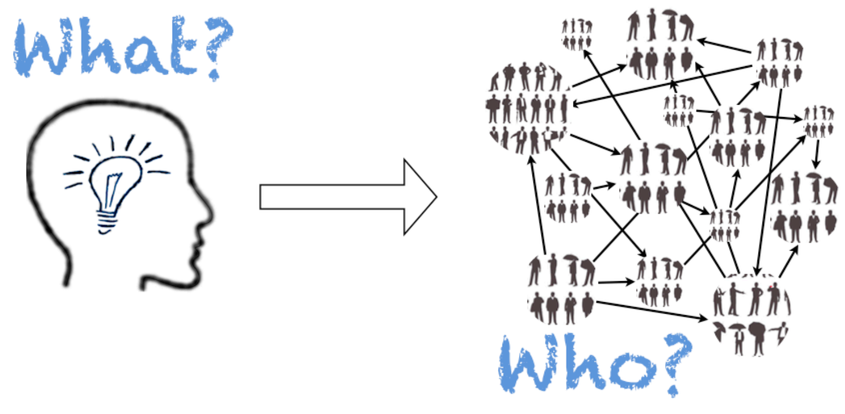
Author Profiling in Social Media: Identifying Information about Gender, Age, Emotions and beyond*

Francisco Manuel Rangel Pardo – @kicorangel
CTO Autoritas / PhD Student at UPV

Strategic Intelligence, from Know-How to Know-Who

"If you know the enemy and know yourself, you need not fear the result of hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat. If you know neither the enemy nor yourself, you will sucumb in every battle"

Sun Tzu



Demographics

- ✓ Organization of PAN-AP 2013 Task at Valencia
  http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-profiling.html
- ✓ Objective -> Identify Age & Gender
  - ✓ 10s (13-17), 20s (23-27), 30s (33-47)
  - ✓ Male / female
  - ✓ Two languages (EN / ES)
- ✓ Dataset -> Social media, some challenges
  - ✓ Large dataset -> big data?
  - ✓ Auto-labeled data
  - ✓ Auto-generated content -> robots, ads...
  - ✓ High variety of themes
  - ✓ Introduction of chatlines from pedophiles

Test Dataset

LANG	AGE	NUM. OF AUTHORS
EN	10s	8,600 / 740 / 888
	20s	42,828 / 3,840 / 4,608
	30s	66,800 / 6,020 / 7,224
ES	10s	1,250 / 120 / 144
	20s	21,300 / 1,920 / 2,304
	30s	15,400 / 1,360 / 1,632

- ✓ Methodology -> Machine learning, high variety of features
  - ✓ Stylometrics, readability, content, lsa...
- ✓ Results ->

Early bird results

TEAM	ENGLISH		SPANISH	
	AGE	GENDER	AGE	GENDER
Gillan	**59.47**	54.13	53.57	47.74
Landra	59.24	**56.31**	57.57	**61.71**
Ayala	2.78	2.77	8.41	8.44
Jankowska	54.63	51.84	44.79	58.34
Baseline	33.24	49.97	33.53	50.01
Rangel	-	-	**62.72**	56.75

- ✓ Conclusions
  - ✓ Difficult task
  - ✓ Gender detection not better than baseline
  - ✓ Our style features are competitive
- ✓ Future work
  - ✓ PAN-AP 2014 is comming...

Emotions

- ✓ Objective -> Identify Eckman's 6 basic emotions
- ✓ Dataset ->
  - ✓ 1,200 comments from Facebook
  - ✓ 3 different themes: politics, football, public people
  - ✓ Manually labeled by 3 annotators

Dataset

	MALE	FEMALE
POLITICS	200	200
FOOTBALL	200	200
PUBLIC PEOPLE	200	200

- ✓ Methodology -> Machine learning, stylistic + dictionary features

FREQUENCIES	Unique words, capital words, capital characters, word length, character flooding
PUNCTUATION	Dots, commas, semicolon, colon, exclamations, questions, quotes
GRAMMATICAL CATEGORIES	Number and person of verbs and pronouns, verb tense, named entities, non-dictionary words
EMOTICONS	Ratio between emoticons and words, numer of each kind of emoticon
SPANISH EMOTION LEXICON*	For each word we obtained the value of the FPA from the dictionary, and all the FPA values for each emotion. * http://www.cic.ipn.mx/~sidorov/#SEL
BoW	Top 20 words with more information gain and independent from the content (adjectives, adverbs...)

- ✓ Results ->

Results*

EMOTION	PREC.	REC.	F
JOY	71.1	68.6	0.695
ANGER	84.5	73.3	0.772
DISGUST	87.3	75.9	0.799
SURPRISE	67.5	67.8	0.676
SADNESS	91.1	80.2	0.845

* No enough results for "fear"

Gender identification accuracy: **53.6%**

- ✓ Conclusions ->
  - ✓ Style features seems to be appropiate to identify emotions
  - ✓ Competitive results compared to SoA (SEMEval07...)
  - ✓ Features also used in PAN-AP task
- ✓ Future work ->
  - ✓ Improve style features -> collocations, speech analysis
  - ✓ Research the relationship between emotions and personality traits

Personality traits and beyond...

Writings of anonymous users is the only thing we can trust...
...and not even that

- ✓ How is emotionality linked to demographics?
- ✓ How are the emotions related to personality traits?
- ✓ How do different personalities express their emotions?
- ✓ How can demographics influence personality?
- ✓ How can the author profiling help us in social network analysis?
- ✓ What are the best features for describing users' style?
- ✓ How may the answer to these questions help us to answer the question "who"?

The main objective is to build a common framework which allow us to better understand how people use language and how such use helps to profile them

autoritas® nuevas ideas, nuevas soluciones

COSMOS

Universitat Politècnica de València

ESSIR 2013 granada.spain