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

Francisco Manuel Rangel Pardo

CTO Autoritas Consulting/
PhD Student at UPV
@kicorangel

autoritas
nuevas ideas, nuevas soluciones

essir 2013
granada.spain

STRATEGIC INTELLIGENCE

"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

autoritas®
nuevas ideas, nuevas soluciones

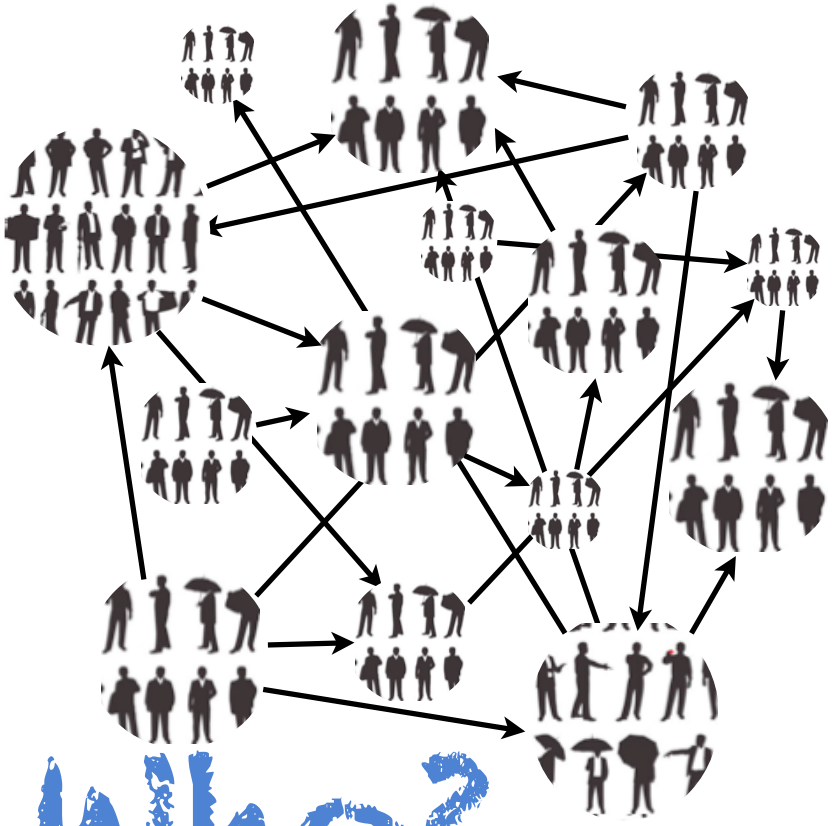
COSMOS

VNIVERSITAT POLITÈCNICA VALÈNCIA

essir 2013
granada.spain

FROM KNOW-HOW TO KNOW-WHO

What?



Who?

autoritas
nuevas ideas, nuevas soluciones

COSMOS

VNIVERSITAT POLITÈCNICA VALÈNCIA

essir 2013 granada.spain

AUTHOR PROFILING - DEMOGRAPHICS

- ✓ Organization of PAN-AP Task at CLEF Valencia 2013
- ✓ Objective -> Identify Age & Gender
  - ✓ 10s (13-17), 20s (23-27), 30s (33-47), Male / female
  - ✓ Two languages (EN / ES)
- ✓ Dataset -> Social media, some issues / challenges
  - ✓ Large dataset -> big data?
  - ✓ Auto-labeled data
  - ✓ Auto-generated content -> robots, ads...
  - ✓ High variety of themes
  - ✓ Introduction of chatlines from pedophiles
- ✓ Methodology -> Machile learning, high variety of features
  - ✓ Stylometrics, readability, content, lsa...
- ✓ Conclusion -> Difficult task
  - ✓ Gender identification not better than baseline

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

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

autoritas
nuevas ideas, nuevas soluciones

COSMOS

VNIVERSITAT POLITÈCNICA VALÈNCIA

2013 essir granada.spain

AUTHOR PROFILING - EMOTIONS

Dataset

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

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

Gender: 53.6%

* No enough results for "fear"

- ✓ Objective -> Identify 6 basic emotions
- ✓ Data ->
  - ✓ 1,200 comments from Facebook
  - ✓ 3 different themes: politics, football, public people
  - ✓ Manually annotated for 6 emotions
- ✓ Features -> Stylistic features + dictionary vs. content features
  - ✓ Punctuation marks such as dots, commas, quotations, question marks and so on, frequencies such as numbers of unique words, capital words, words with character flooding and so on, grammatical categories, verb tenses, verb and pronouns number and person, named entities, non-dictionary words, emoticons and emotion words from the Spanish Emotion Lexicon
  - ✓ Features used in PAN-AP task
- ✓ Conclusion ->
  - ✓ Competitive results compared with SoA
  - ✓ Features valid for demographics identification
- ✓ Future work ->
  - ✓ Improve style features -> collocations
  - ✓ Research the relationship with personality traits

autoritas
nuevas ideas, nuevas soluciones

COSMOS

VNIVERSITAT POLITÈCNICA VALÈNCIA

2013 essir granada.spain

PERSONALITY TRAITS AND BEYOND...

The writings of anonymous users is the only thing we can trust...

...and not even that

- ✓ How is emotionality linked to demographics?
- ✓ How are emotions related to personality traits?
- ✓ How do different personalities express their emotions?
- ✓ How does demographics influence in 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 allows us to better understand how people use language and how such use helps to profile them

autoritas
nuevas ideas, nuevas soluciones

COSMOS

VNIVERSITAT POLITÈCNICA VALÈNCIA

2013 essir granada.spain