



CIIT/inIT, Lemgo, 13th January 2014

AUTORITAS

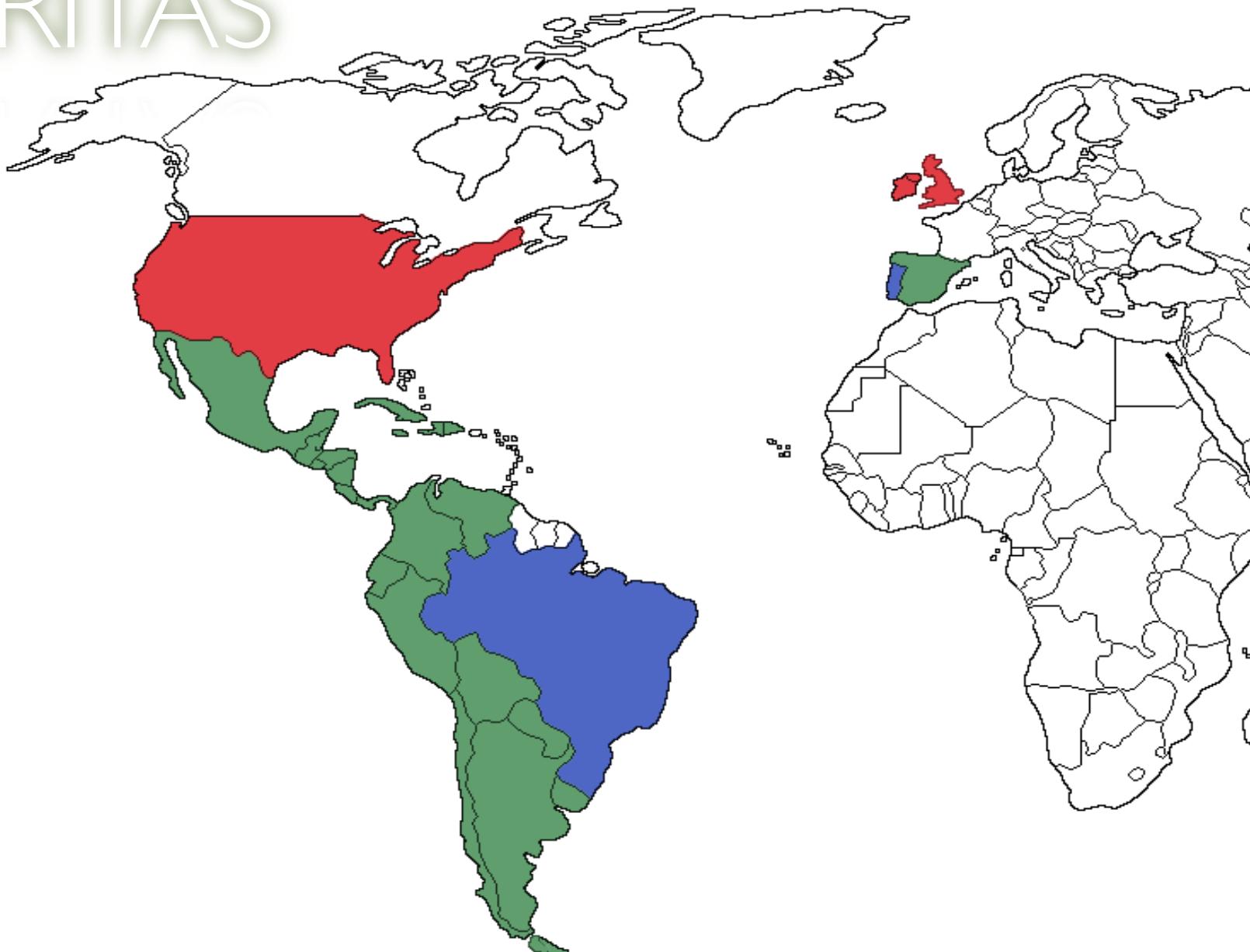
Autoritas is a technological consultancy with multidisciplinary professionals from many different countries of the world:

- Computer scientists
- Psychologists
- Economists
- Lawyers
- Political Scientists



AUTORITAS

Our headquarter is in Madrid (Spain), and we have offices in Argentina, Brazil, Chile, Mexico, Panama and UK.



AUTORITAS

We specialize in:

- Spanish
- Portuguese
- English

We are starting with:

- Italian
- German
- French



AUTORITAS

Training and communication are very important for us.

Active Listening is a new discipline and we should train professionals interested on the field.

autoritas®

nuevas ideas, nuevas soluciones



Socialancer.com



Iberolabs

AUTORITAS

We make an active research in collaboration with reputed universities and research centers.

- ▶ **2008-2009: Avanza AL4Tour (ITC/464/2008)**

Objective: Develop an Active Listening platform specialized in the Tourism Sector

Budget: 744,681.83 €

- ▶ **2011-2012: Avanza Competitividad HCE (TSI-020100-2011-156)**

Objective: Specialize the Active Listening platform in the Healthcare Sector by developing an automatic coder of diseases expressed in natural language into ICD-9 MC codes.

Budget: 432,129.25€

- ▶ **2012: InnImpacto Ecoportunity (IPT-2012-1220-430000)**

Objective: Combine macroeconomic indicators with social media information to predict trends.

Budget: 576,629.00€

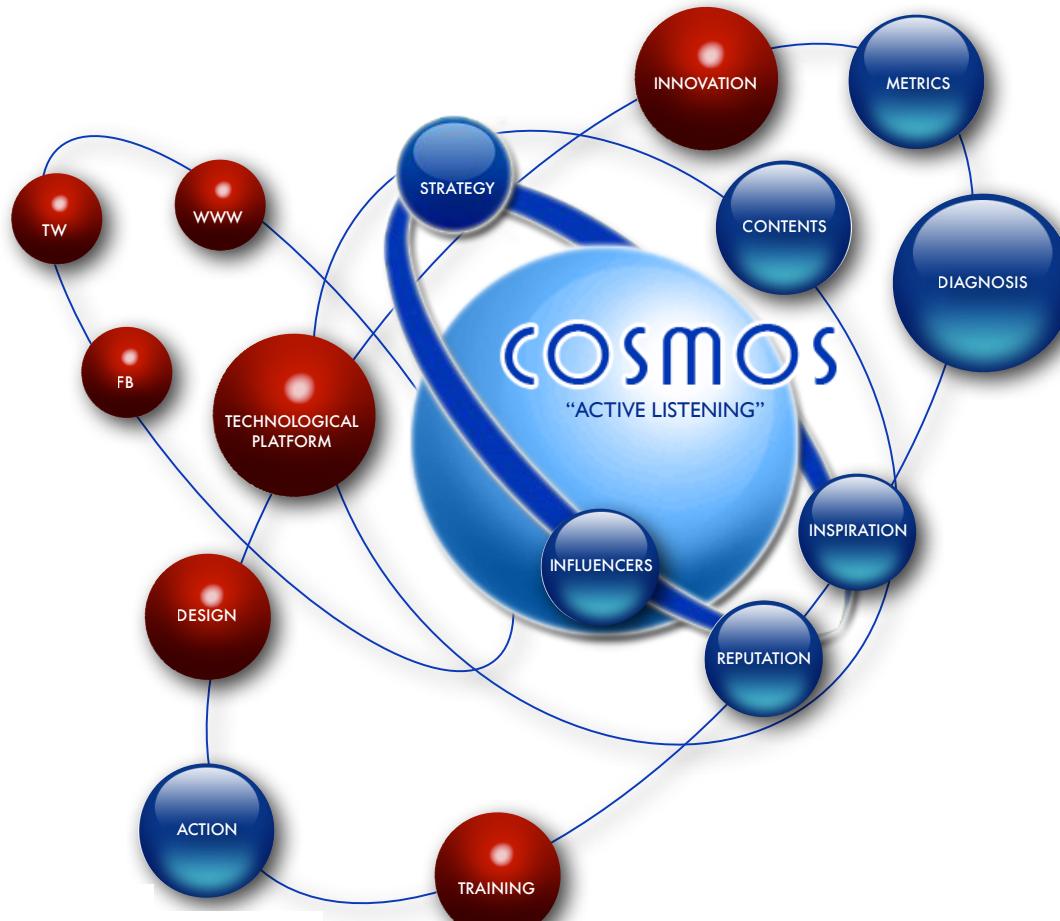
AUTORITAS

Our main objective is to help companies to interact with the universe of Internet in order to incorporate Internet in their decision making process.

autoritas®
nuevas ideas, nuevas soluciones

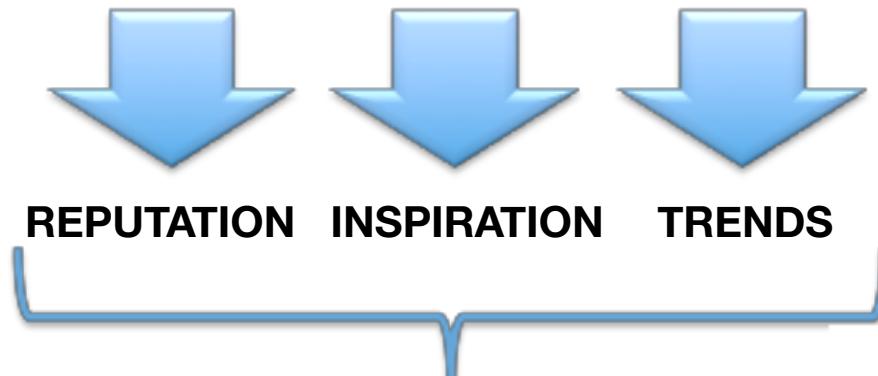


What is Cosmos?



- ▶ **Cosmos is an active listening tool developed by Autoritas Consulting.**
“Active Listening is more than paying attention” Margaret Mead
- ▶ **Cosmos brings order to the chaos of the internet and provides a methodology for individuals and workteams who want to make the most of the opportunities that arise.**
- ▶ **Cosmos aims to help organizations touch the internet and get knowledge for the decision making process.**

Active Listening is BSI



- ▶ Cosmos deals with the entire internet universe, covering all the needs an institution might have.
 - ▶ Where can we get some inspiration?
 - ▶ What are people saying about us?
 - ▶ What impact do we have?
 - ▶ What is our target interested in?



Active Listening Cycle



Knowledge Extraction

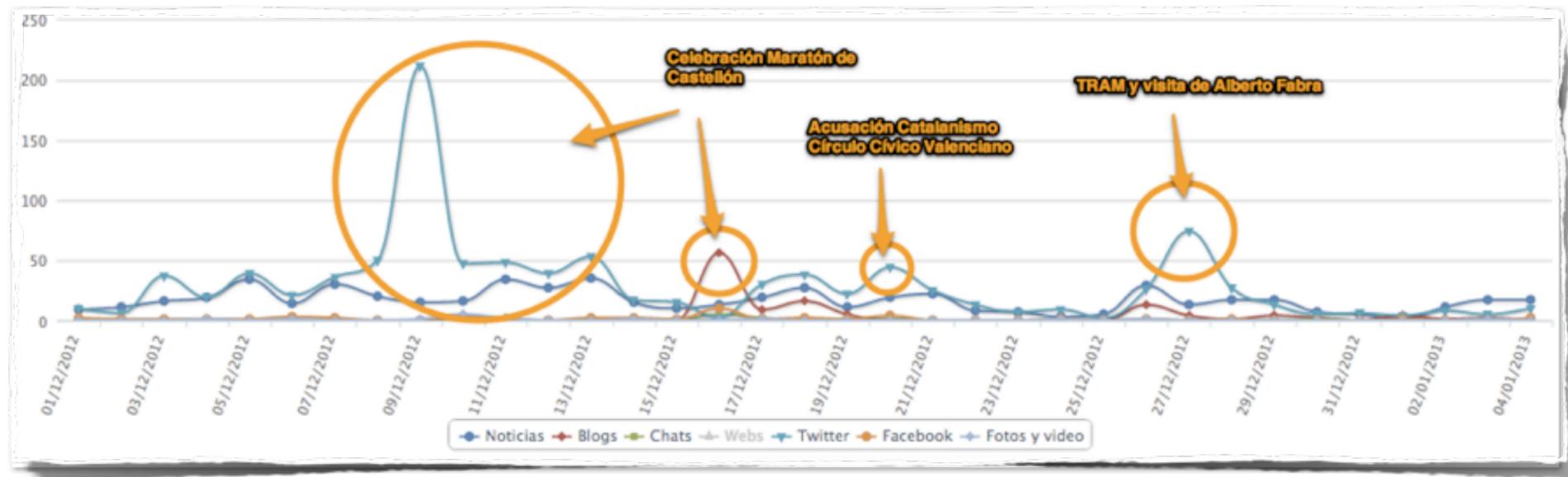


What? -> Conceptualizing brand



What are people talking about?

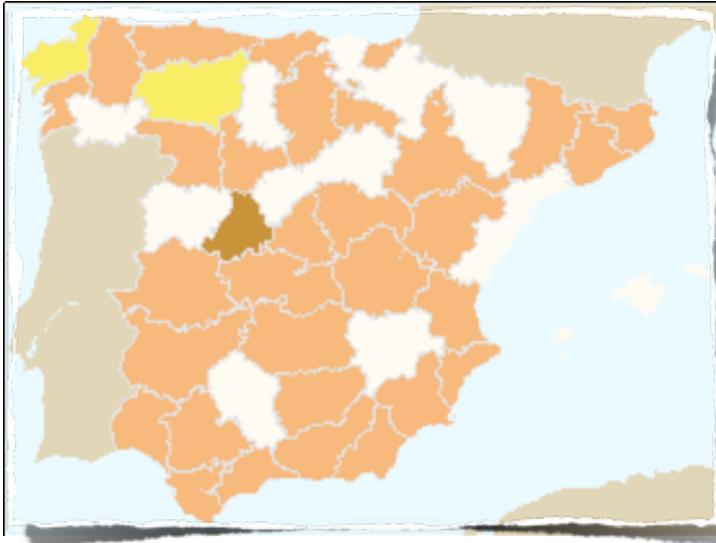
When? -> Crisis management



When is something happening?



Where? -> In two dimensions



Where does a conversation occur?

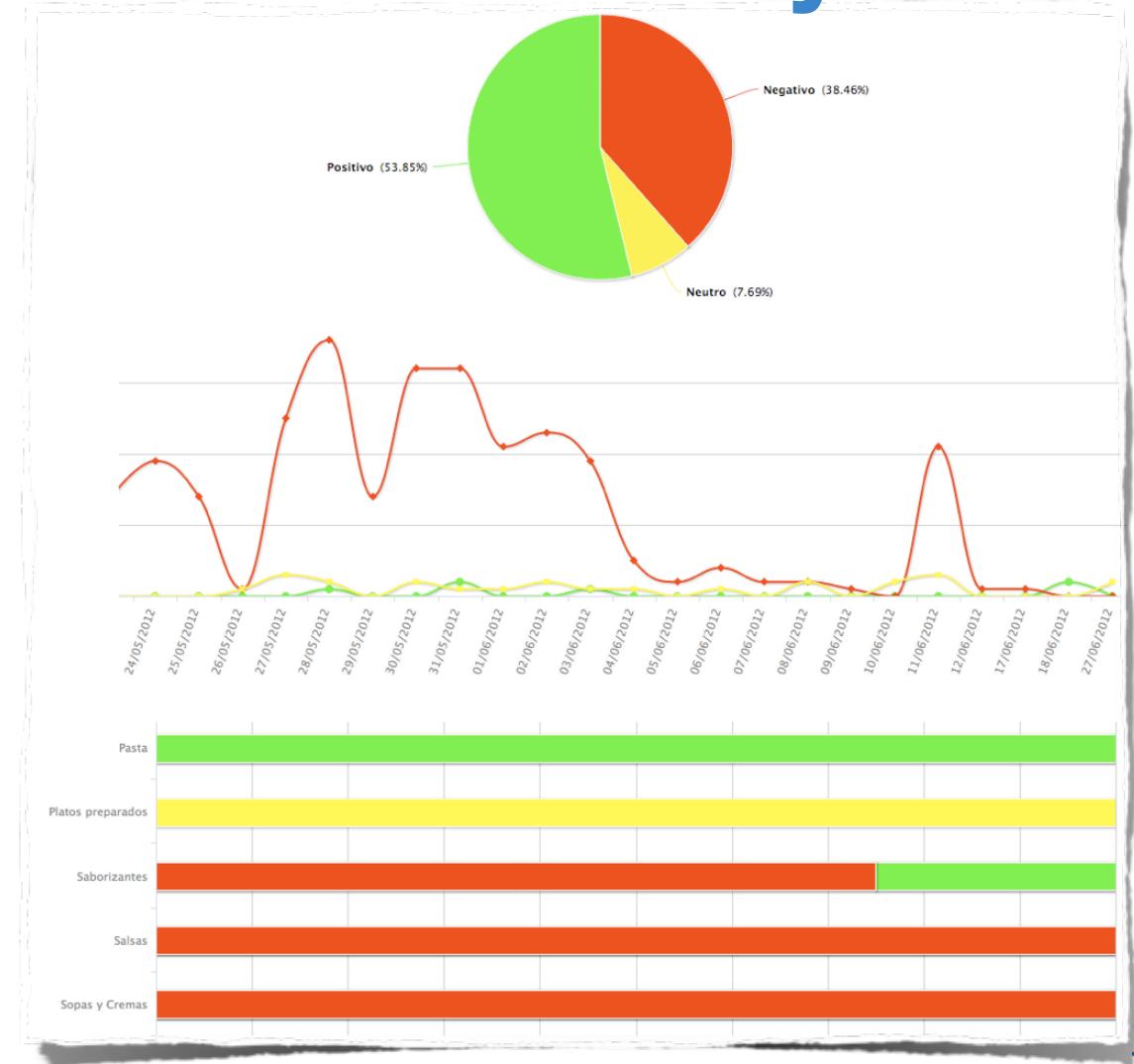


Where is the focus of a conversation?



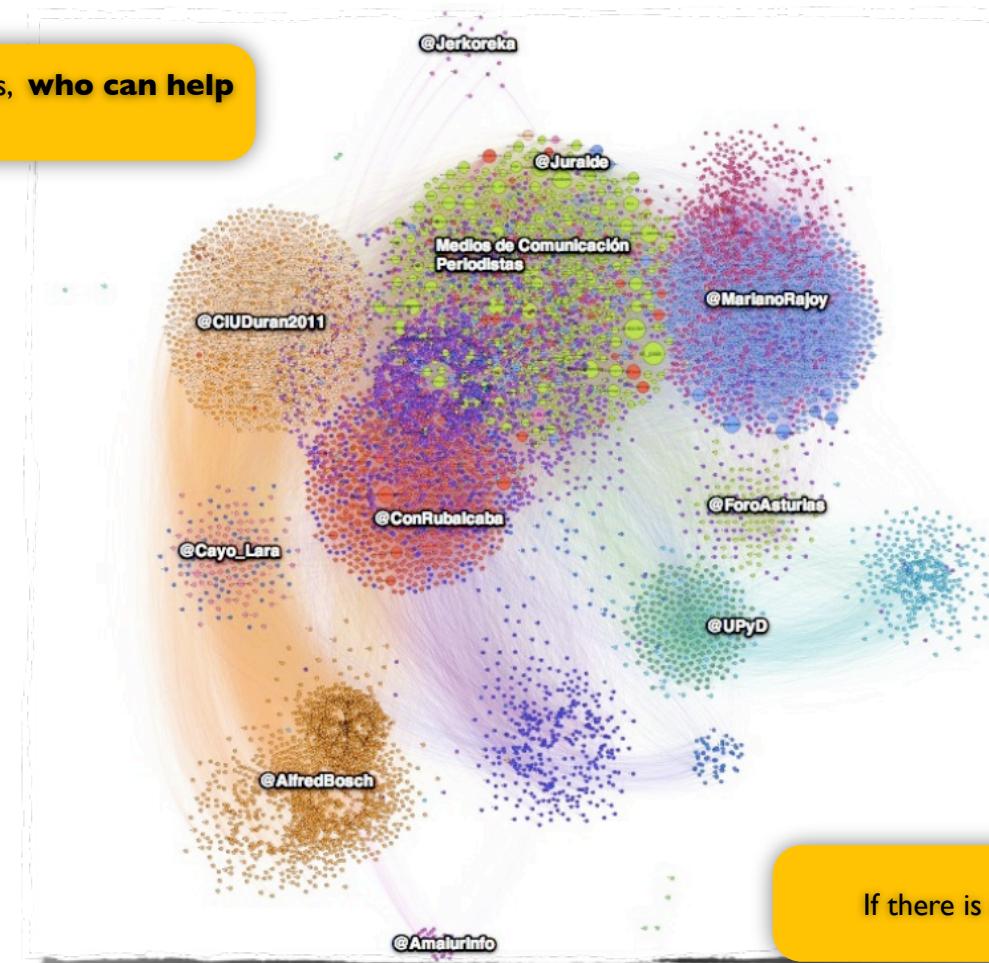
How? -> Not only Sentiment Analysis

Polarity is only one dimension, emotional, values, SWOT... all of them answer to HOW



Who? -> Social Network Analysis

If I want to transmit a message with success, **who can help me?**

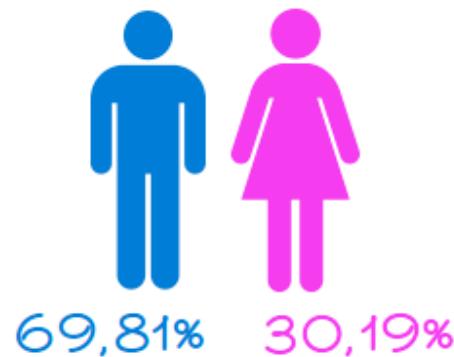


If there is a conflict, **who do I have to watch?**

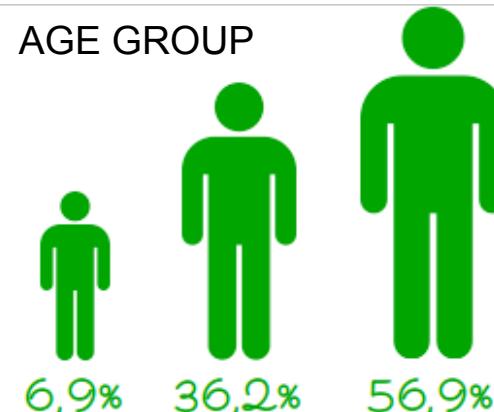


Why? -> Author Profiling

GENDER



AGE GROUP

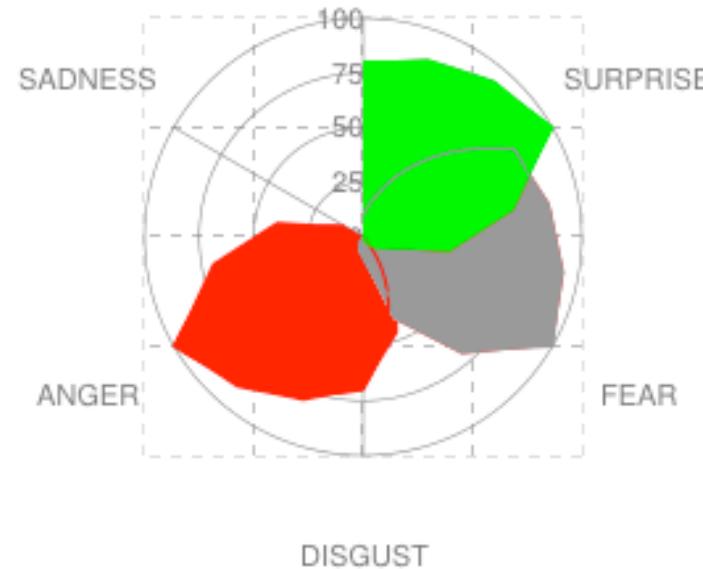


NATIVE LANGUAGE



EMOTIONAL PROFILE

JOY



PERSONALITY TRAITS





In collaboration with:
Paolo Rosso



UNIVERSITAT
POLITÈCNICA
DE VALENCIA

Language | Langue | Linguaggio
Языка | Sprache | SPRACHE
Lingua | LLENGUAJE | لغة
NTEL
Natural Language Engineering Lab

PAN-AP at CLEF 2013

- ✓ Objective
 - ✓ Identify Age and Gender
 - ✓ Identify age and gender of sexual predators/pedophiles
 - ✓ Measure computational time needed to perform the task
- ✓ 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
- ✓ Labeling
 - ✓ 10s (13-17), 20s (23-27), 30s (33-47), Male / female
 - ✓ Two languages (EN / ES)
- ✓ Methodology -> Machine learning, high variety of features
 - ✓ Stylometrics, readability, content, Is...
- ✓ Conclusion -> Difficult task
 - ✓ Gender identification not better than baseline
 - ✓ Even more difficult for gender & age together
 - ✓ Robust to detect predators
 - ✓ Expensive in time consuming



Test Dataset

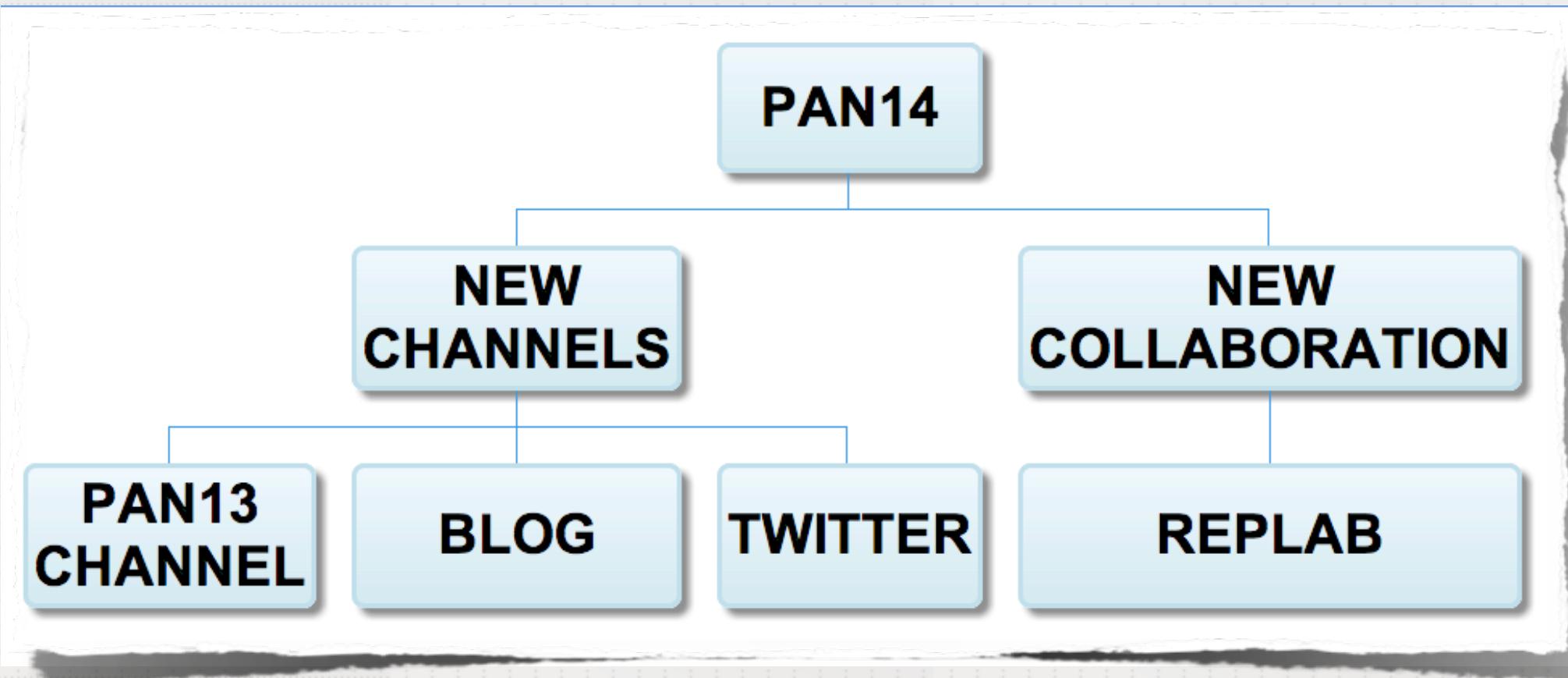
LANG	AGE	NUM. OF AUTHORS
EN	10s	8,600 / 740 / 888
	20s	42,828 / 3,840 / 4,608
	30s	66,800 / 6,020 / 7,224
ES	10s	1,250 / 120 / 144
	20s	21,300 / 1,920 / 2,304
	30s	15,400 / 1,360 / 1,632

Top 10 results

English				Spanish			
Team	Total	Gender	Age	Team	Total	Gender	Age
Meina	0.3894	0.5921	0.6491	Santosh	0.4208	0.6473	0.6430
Pastor L.	0.3813	0.5690	0.6572	Pastor L.	0.4158	0.6299	0.6558
Seifeddine	0.3677	0.5816	0.5897	Cruz	0.3897	0.6165	0.6219
Santosh	0.3508	0.5652	0.6408	Flekova	0.3683	0.6103	0.5966
Yong Lim	0.3488	0.5671	0.6098	Ladra	0.3523	0.6138	0.5727
Ladra	0.3420	0.5608	0.6118	De-Arteaga	0.3145	0.5627	0.5429
Aleman	0.3292	0.5522	0.5923	Kern	0.3134	0.5706	0.5375
Gillam	0.3268	0.5410	0.6031	Yong Lim	0.3120	0.5468	0.5705
Kern	0.3115	0.5267	0.5690	Sapkota	0.2934	0.5116	0.5651
Cruz	0.3114	0.5456	0.5966	Pavan	0.2824	0.5000	0.5643

*A lot of interest (66 registered teams)
and a lot of participants (21)*

PAN-AP at CLEF 2014



Emotions and Author Profiling

Results

Emotion	Algorithm	r	Prec.	Rec.	F1
Joy	J48	27.1	49.7	43.2	46.2
	NB	27.9	45.4	56.8	50.5
	BN	25.6	40.9	73.7	52.6
	SVM	24.9	56.9	30.5	39.7
Anger	J48	16.6	32.3	19.9	24.6
	NB	22.6	25.9	60.3	36.3
	BN	22.2	25.6	60.9	36.0
	SVM	10.8	25.8	15.2	19.2
Disgust	J48	21.7	36.1	23.3	28.3
	NB	15.7	19.7	55.8	29.1
	BN	24.9	25.5	64.3	36.5
	SVM	6.2	11.7	5.4	7.4
Surprise	J48	25.8	50.4	48.7	49.5
	NB	20.6	42.7	67.2	52.2
	BN	20.7	43.0	64.6	51.6
	SVM	17.2	49.4	30.5	37.7
Sadness	J48	12.1	20.0	14.5	16.8
	NB	6.1	9.8	35.5	15.4
	BN	16.7	16.3	51.3	24.7
	SVM	8.2	17.9	0.92	12.2
Average results					
J48 20.7 37.7 29.9 33.1					
NB 18.6 28.7 55.1 36.7					
BN 22.0 30.3 63.0 40.3					
SVM 13.5 32.3 16.5 23.2					

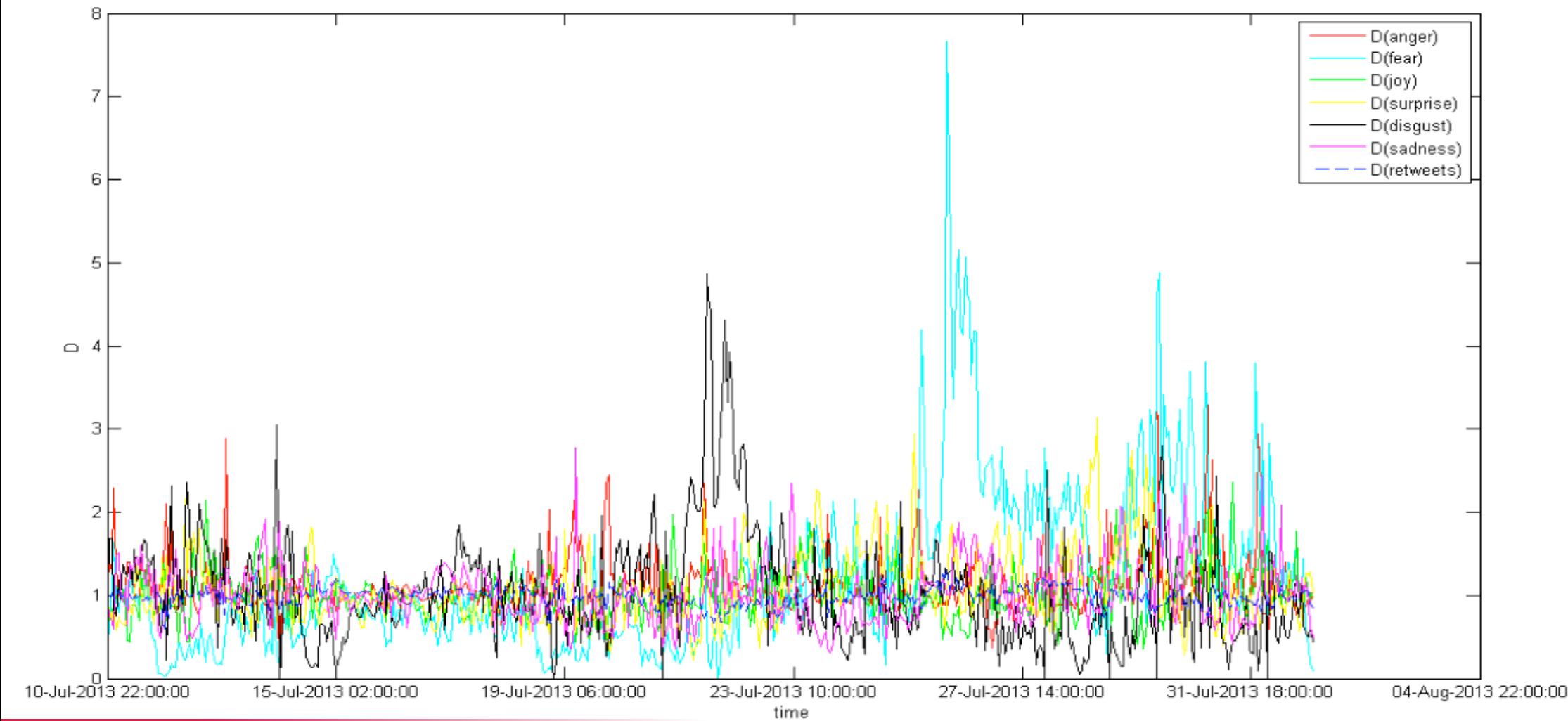
Gender: 59.0%



* No enough results for "fear"

- ✓ Objective -> Identify 6 basic emotions + Gender
- ✓ 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
 - ✓ Emotions for Author Profiling
 - ✓ Research the relationship with personality traits

Emotions for trends identification



what's ?
next



FRANCISCO RANGEL

CTO at AUTORITAS CONSULTING
francisco.rangel@autoritas.es