Overview of the 7th PAN author profiling shared task on:

Bots and gender profiling

Francisco Rangel & Paolo Rosso

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

CLEF 2019 - Lugano

10th September 2019

Bots: propaganda, fake news, inflammatory content

- Bots may **influence** users with **comercial, political or ideological** purposes…

- **Polarization** and spread **disinformation and fake news**

- US 2016 Presidencial election, Brexit, 1 Oct 2017 referendum for the Catalan independence:

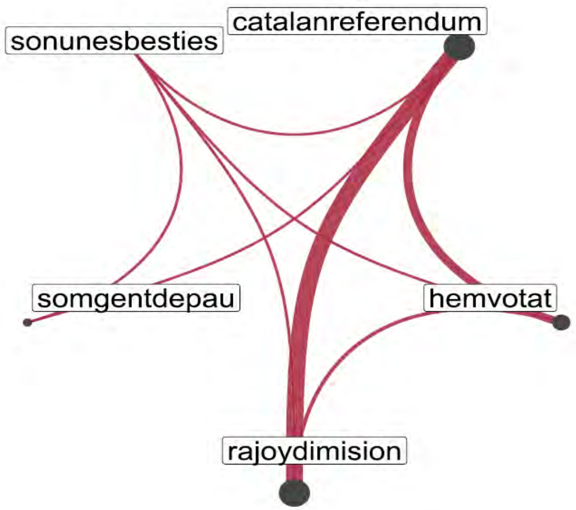
Bots: propaganda, fake news, inflammatory content

- Bots may **influence** users with **comercial, political or ideological** purposes…

- **Polarization** and spread **disinformation and fake news**

- US 2016 Presidencial election, Brexit, 1 Oct 2017 referendum for the Catalan independence:

  23.5% of 3.6 million tweets generated **by bots**
  19% of the interactions were **from bots to humans**

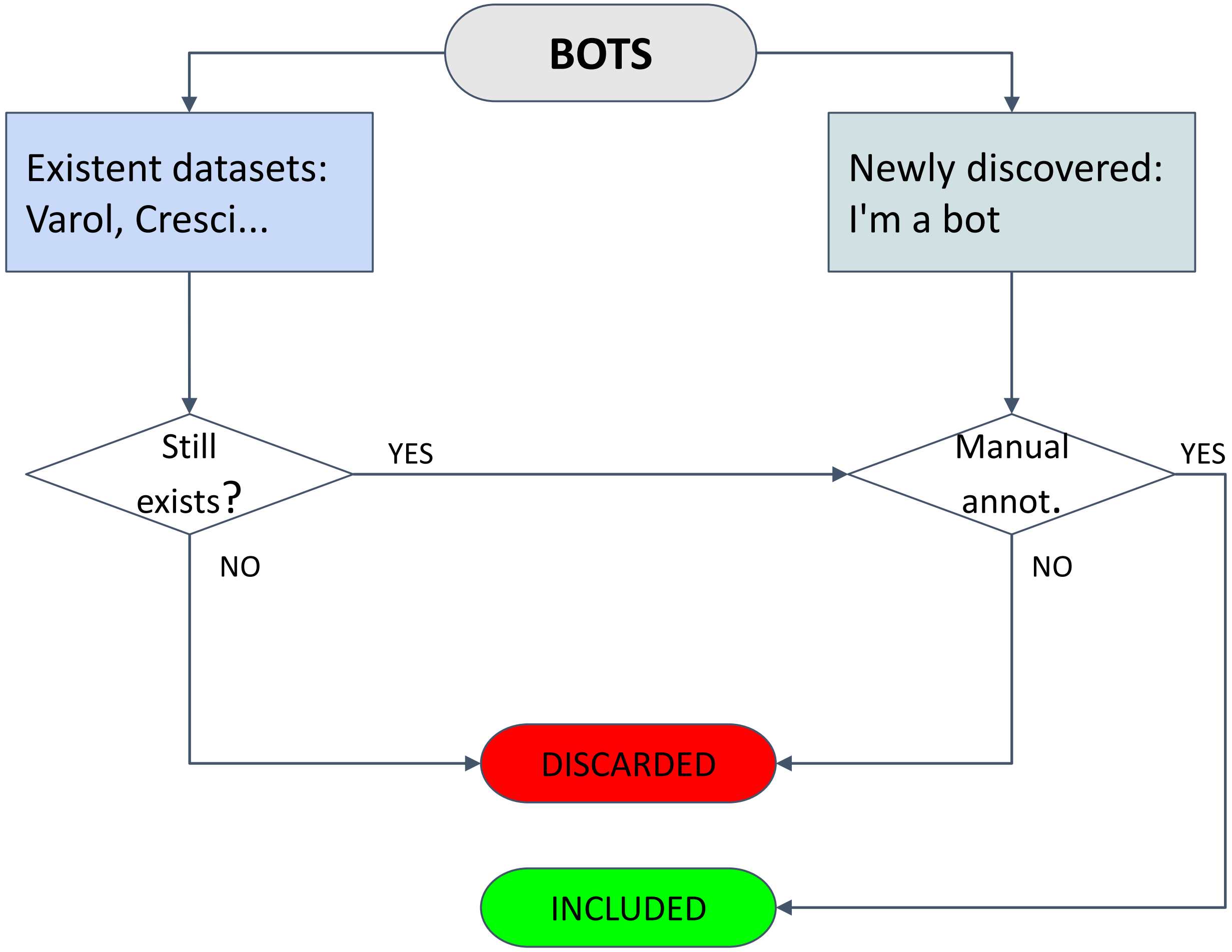


Massimo Stella, Emilio Ferrara, and Manlio De Domenico. Bots increase exposure to negative and inflammatory content in online social systems. Proc. of the National Academy of Sciences of the United States of America, 115(49):12435–12440, 2018.

Bots and gender profiling

- How difficult / easy is to discriminate **bots from humans** on the basis only on **textual features**?
- What are the **most difficult type of bots**?

Bots and humans accounts



Humans selected from PAN-AP'17 author profiling+ manual annotation

Dataset

- Twitter accounts identified as bots in existing datasets + new ones
- Each **author (bot or human) feed** is composed by exactly **100 tweets**

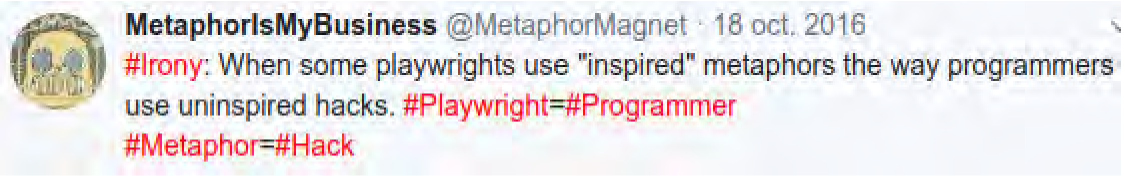
		(EN) English				(ES) Spanish			
		Bots	Humans		Total	Bots	Humans		Total
			F	M			F	M	
Training	**Training**	1,440	720	720	2,880	1,040	520	520	2,080
	Development	620	310	310	1,240	460	230	230	920
	Total	2,060	1,030	1,030	4,120	1,500	750	750	3,000
Test		1,320	660	660	2,640	900	450	450	1,800
Total		3,380	1,690	1,690	6,760	2,400	1,200	1,200	4,800

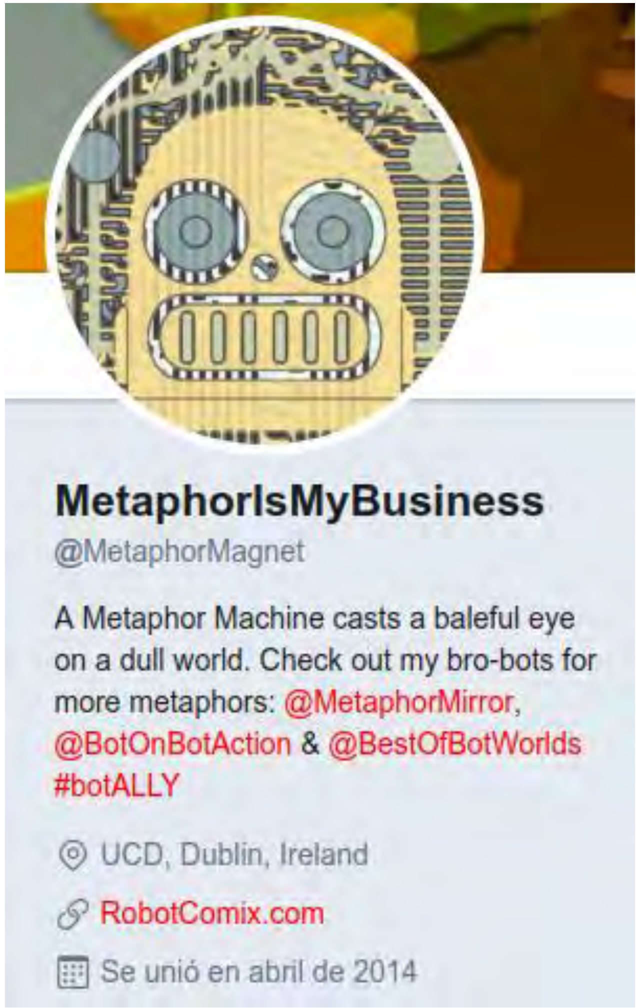
Types of bots

TEMPLATE	The Twitter feed responds to a **predefined structure or template**, such as for example a Twitter account giving the state of the earthquakes in a region or **job offers** in a sector
FEED	The Twitter feed retweets or **shares news about a predefined topic**, such as for example regarding Trump's policies
QUOTE	The Twitter feed reproduces **quotes from famous books or songs, from celebrities** or people, or jokes
ADVANCED	Twitter feeds whose **language is generated** on the basis of more elaborated technologies such as Markov chains, **metaphors**, or in some cases, randomly choosing and merging texts from big corpora

Metaphormagnet

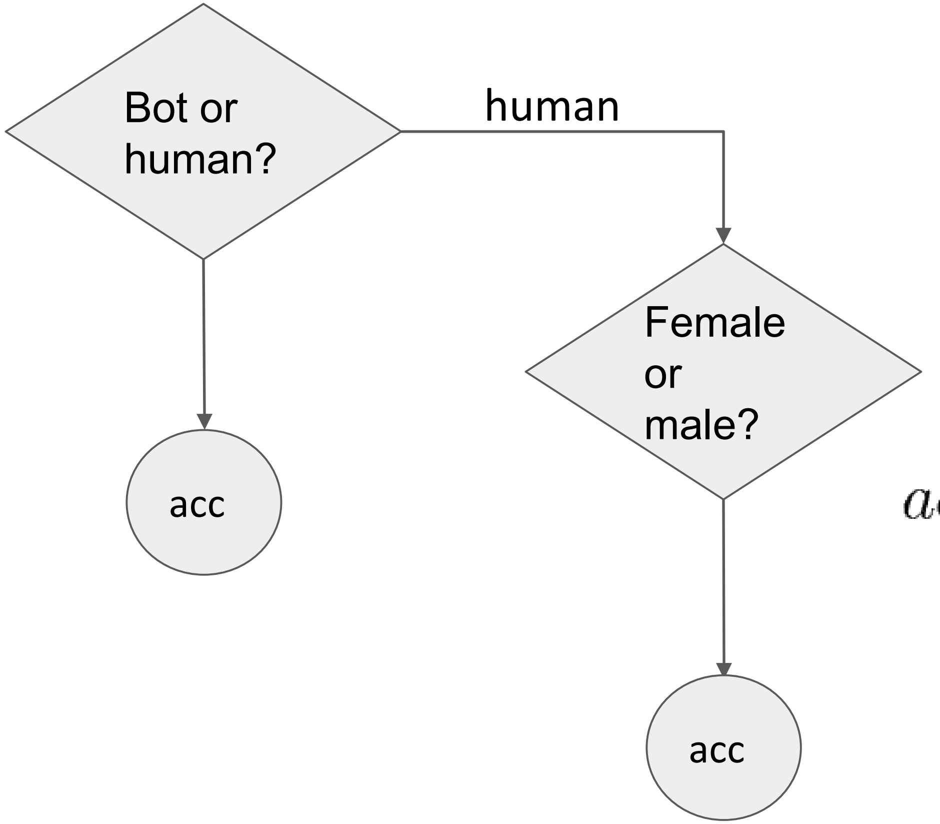
For example, the bot **@metaphormagnet** was developed by **Tony Veale** and **Goufu Li** to automatically generate metaphorical language





Evaluation measures

Accuracy is calculated per language and task:



$$acc_{[en|es]} = \frac{acc_{bots} + acc_{gender}}{2}$$

$$ranking = \frac{acc_{en} + acc_{es}}{2}$$

Statistics

55+1 participants
26 countries

Approaches

What kind of ...

Preprocessing | Features | Methods

... did the teams perform?

Approaches: Preprocessing

Twitter elements (URLs, users, hashtags, ...)	Van Halteren; Vogel; Polignano; Giachanou; Gishamer; Puertas; Saeed; Petritk; Valencia; Onose; Babaei; Yacob; Zhechev; Mahmood
Word segmentation	Gishamer; Joo
Tokenisation	Van Halteren; Polignano; Gishamer; Joo; Bacciu; Petritk; Goubin; Zhechev; Mahmood
Stemming / lemmatisation	Ikae; Joo; Saeed; Bacciu; Basile; Petritk; Babaei; Goubin; Zhechev;
Punctuation marks	Vogel; Saeed; Onose; Ribeiro; Goubin; Yacob; Zhechev;
Lowercase	Van Halteren; Vogel; Giachanou; Saeed; Ribeiro
Stopwords	Joo; Saeed; Babaei; Zhechev;
Character flooding	Vogel; Gishamer; Goubin
Latent Semantic Analysis	Rakesh
Short words	Vogel
Infrequent words	Ikae; Gishamer
Contractions and acronyms	Joo; Saeed

Approaches: Features

Stylistic features: - Number of occurrences - Verbs, adjs, pronouns - Number of hashtags, mentions, URLs... - Upper vs. lower case - Punctuation marks - ...	Joo; Goubin; Ashraf; Cimino; Oliveira; Ikae; De la Peña; Johansson; Giachanou; Martinc; Przybyla; Van Halteren; Fernquist
N-gram models	Ispas; Bounaama; Rakesh; Valencia; Mahmood; Fahim; Espinosa; Pizarro; Martinc; Martinc; Dias; Vogel; Giachanou; De la Peña; Babaei; Saeed; Joo; Bacciu; Johansson; Fernquist; HaCohen; Gishamer
Emotional features	Cimino; Giachanou; Oliveira
Lexicon-based features	Gamallo
Compression algorithms	Fernquist
DNA-based approach	Kosmajac
Embeddings	Polignano; Fagni; Halvani; Onose; López-Santillán; Staykovsky; Joo

Approaches: Methods

SVM	Vogel; Cimino; Fagni; Pizarro; Jimenez; HaCohen; Bacciu; Goubin; Srinivasarao; Mahmood; Yacob; Ribeiro; Babaei; Rakesh; Gishamer; Moryossef; Giachanou		
Logistic regression	Gishamer; Moryossef; Valencia; Bolonyai; Przybyła	CatBoost	Fernquist
SpaCy	Moryossef	kNN	Ikae
Random Forest	Moryossef; Johansson	Multilayer Perceptron	Staykovski
Stochastic Gradient Descent	Giachanou; Bounaama	RNN	Dias; Petrik; Bolonyai; Onose
Decision Trees	Saeed	CNN	Dias; Petrik; Polignano; Farber
Multinomial BayesNet	Saeed	BERT	Joo
Naive Bayes	Gamallo	Feedforward NN	Halvani; De la Peña
Adaboost	Bacciu	LSTM	Zhechev

Baselines

MAJORITY	A statistical baseline that always predicts the majority class in the training set. In case of balanced classes, it predicts one of them
RANDOM	A baseline that randomly generates the predictions among the different classes
CHAR N-GRAMS	With values for n from 1 to 10, and selecting the 100, 200, 500, 1,000, 2,000, 5,000 and 10,000 most frequent ones
WORD N-GRAMS	With values for n from 1 to 10, and selecting the 100, 200, 500, 1,000, 2,000, 5,000 and 10,000 most frequent ones
W2V	Texts are represented with two word embedding models: Continuous Bag of Words (CBOW); and Skip-Grams
LDSE	This method represents documents on the basis of the probability distribution of occurrence of their words in the different classes. The key concept of LDSE is a weight, representing the probability of a term to belong to one of the different categories: human / bot, male / female. The distribution of weights for a given document should be closer to the weights of its corresponding category. LDSE takes advantage of the whole vocabulary

Global ranking

Ranking	Team	Bots vs. Human		Gender		Average
		EN	ES	EN	ES	
1	Pizarro	0.9360	**0.9333**	0.8356	**0.8172**	**0.8805**
2	Srinivasarao & Manu	0.9371	0.9061	0.8398	0.7967	0.8699
3	Bacciu et al.	0.9432	0.9078	0.8417	0.7761	0.8672
4	Jimenez-Villar et al.	0.9114	0.9211	0.8212	0.8100	0.8659
5	Fernquist	0.9496	0.9061	0.8273	0.7667	0.8624
6	Mahmood	0.9121	0.9167	0.8163	0.7950	0.8600
7	Ipsas & Popescu	0.9345	0.8950	0.8265	0.7822	0.8596
8	Vogel & Jiang	0.9201	0.9056	0.8167	0.7756	0.8545
9	Johansson & Isbister	**0.9595**	0.8817	0.8379	0.7278	0.8517
10	Goubin et al.	0.9034	0.8678	0.8333	0.7917	0.8491
11	Polignano & de Pinto	0.9182	0.9156	0.7973	0.7417	0.8432
12	Valencia et al.	0.9061	0.8606	**0.8432**	0.7539	0.8410
13	Kosmajac & Keselj	0.9216	0.8956	0.7928	0.7494	0.8399
14	Fagni & Tesconi	0.9148	0.9144	0.7670	0.7589	0.8388
	char nGrams	0.9360	0.8972	0.7920	0.7289	0.8385
15	Glocker	0.9091	0.8767	0.8114	0.7467	0.8360
	word nGrams	0.9356	0.8833	0.7989	0.7244	0.8356
16	Martinc et al.	0.8939	0.8744	0.7989	0.7572	0.8311
17	Sanchis & Velez	0.9129	0.8756	0.8061	0.7233	0.8295
18	Halvani & Marquardt	0.9159	0.8239	0.8273	0.7378	0.8262
19	Ashraf et al.	0.9227	0.8839	0.7583	0.7261	0.8228
20	Gishamer	0.9352	0.7922	0.8402	0.7122	0.8200
21	Petrik & Chuda	0.9008	0.8689	0.7758	0.7250	0.8176
22	Oliveira et al.	0.9057	0.8767	0.7686	0.7150	0.8165
	W2V	0.9030	0.8444	0.7879	0.7156	0.8127
23	De La Peña & Prieto	0.9045	0.8578	0.7898	0.6967	0.8122
24	López Santillán et al.	0.8867	0.8544	0.7773	0.7100	0.8071
	LDSE	0.9054	0.8372	0.7800	0.6900	0.8032
25	Bolonyai et al.	0.9136	0.8389	0.7572	0.6956	0.8013

Global ranking

26	Moryossef	0.8909	0.8378	0.7871	0.6894	0.8013
27	Zhechev	0.8652	0.8706	0.7360	0.7178	0.7974
28	Giachanou & Ghanem	0.9057	0.8556	0.7731	0.6478	0.7956
29	Espinosa et al.	0.8413	0.7683	0.8413	0.7178	0.7922
30	Rahgouy et al.	0.8621	0.8378	0.7636	0.7022	0.7914
31	Onose et al.	0.8943	0.8483	0.7485	0.6711	0.7906
32	Przybyla	0.9155	0.8844	0.6898	0.6533	0.7858
33	Puertas et al.	0.8807	0.8061	0.7610	0.6944	0.7856
34	Van Halteren	0.8962	0.8283	0.7420	0.6728	0.7848
35	Gamallo & Almatarneh	0.8148	0.8767	0.7220	0.7056	0.7798
36	Bryan & Philipp	0.8689	0.7883	0.6455	0.6056	0.7271
37	Dias & Paraboni	0.8409	0.8211	0.5807	0.6467	0.7224
38	Oliva & Masanet	0.9114	0.9111	0.4462	0.4589	0.6819
39	Hacohen-Kerner et al.	0.4163	0.4744	0.7489	0.7378	0.5944
40	Kloppenburg	0.5830	0.5389	0.4678	0.4483	0.5095
	MAJORITY	0.5000	0.5000	0.5000	0.5000	0.5000
	RANDOM	0.4905	0.4861	0.3716	0.3700	0.4296
41	Bounaama & Amine	0.5008	0.5050	0.2511	0.2567	0.3784
42	Joo & Hwang	0.9333	-	0.8360	-	0.4423
43	Staykovski	0.9186	-	0.8174	-	0.4340
44	Cimino & Dell'Orletta	0.9083	-	0.7898	-	0.4245
45	Ikae et al.	0.9125	-	0.7371	-	0.4124
46	Jeanneau	0.8924	-	0.7451	-	0.4094
47	Zhang	0.8977	-	0.7197	-	0.4044
48	Fahim et al.	0.8629	-	0.6837	-	0.3867
49	Saborit	-	0.8100	-	0.6567	0.3667
50	Saeed & Shirazi	0.7951	-	0.5655	-	0.3402
51	Radarapu	0.7242	-	0.4951	-	0.3048
52	Bennani-Smires	0.9159	-	-	-	0.2290
53	Gupta	0.5007	-	0.4044	-	0.2263
54	Qurdina	0.9034	-	-	-	0.2259
55	Aroyehun	0.5000	-	-	-	0.1250

Best results

Johansson
- Stylistic features
- Random Forest

Valencia
- n-grams
- Logistic Regression

Language	Bots vs. Human	Gender
English	0.9595	0.8417
Spanish	0.9333	0.8172

Pizarro
- n-grams
- SVM

Confusion matrices: bots vs. humans

English

Spanish

Confusion matrices: gender

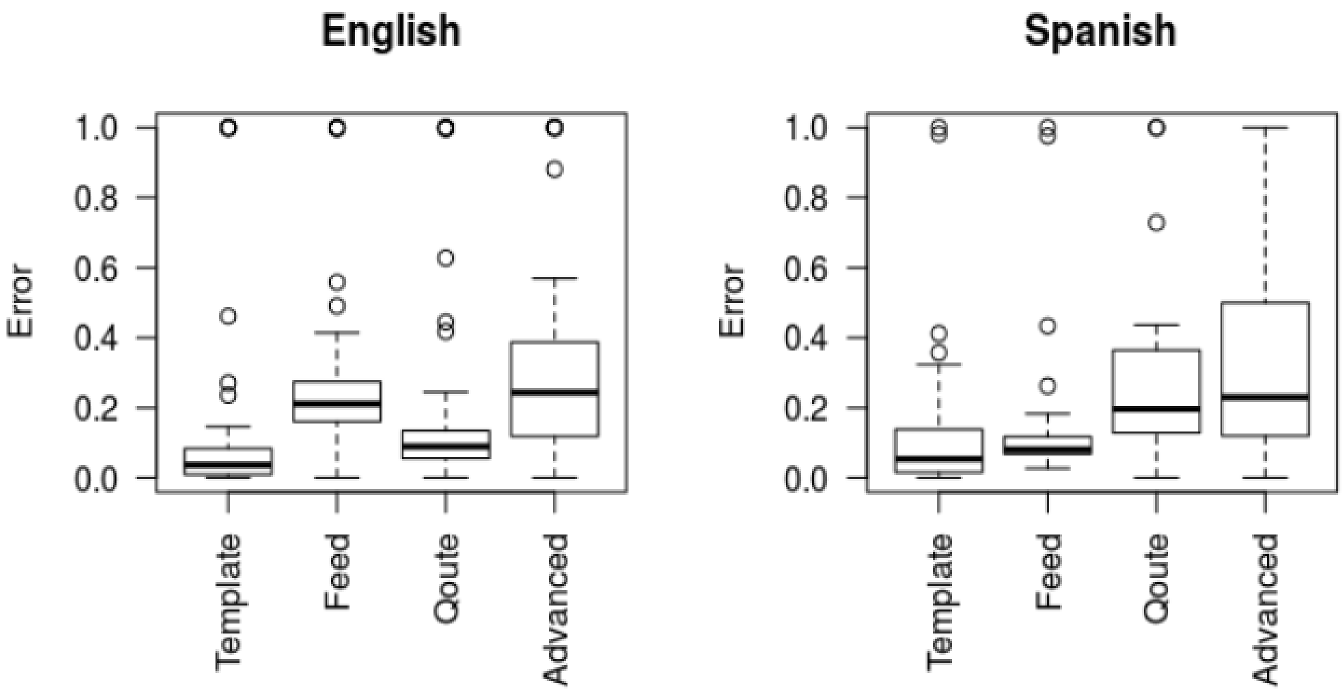
English

	Bot	Male	Female
Female	0.0355	0.2667	0.6978
Male	0.0885	0.6359	0.2756
Bot	0.8264	0.0983	0.0753

Spanish

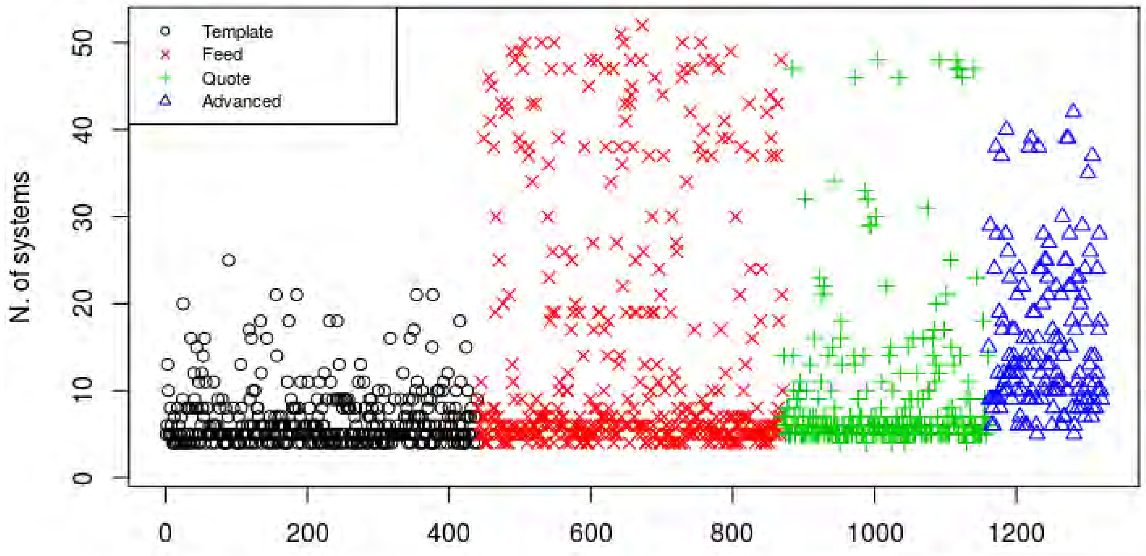
	Bot	Male	Female
Female	0.1161	0.3634	0.5205
Male	0.1893	0.6004	0.2103
Bot	0.8648	0.085	0.0502

Errors per bot type

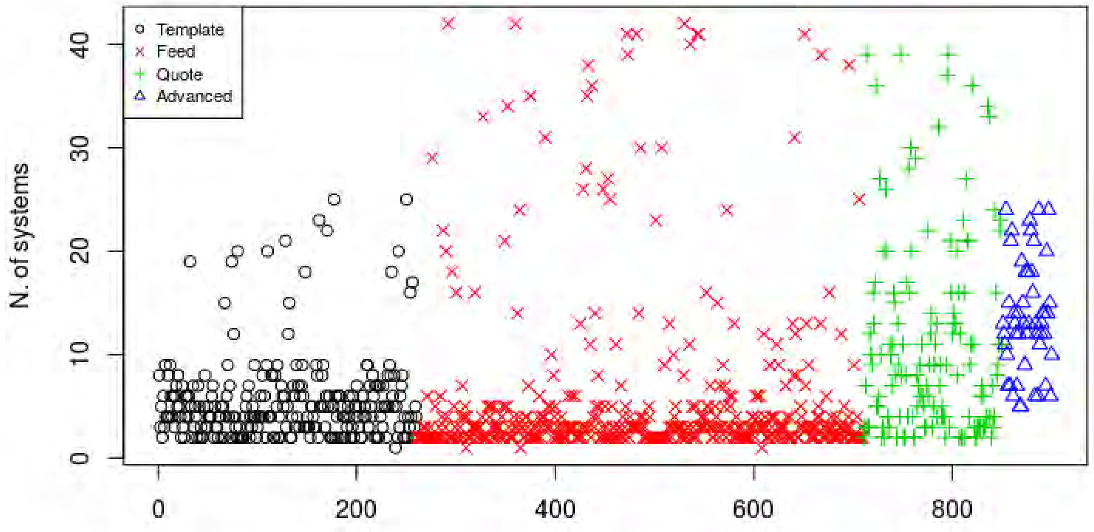


Errors per bot type

ENGLISH



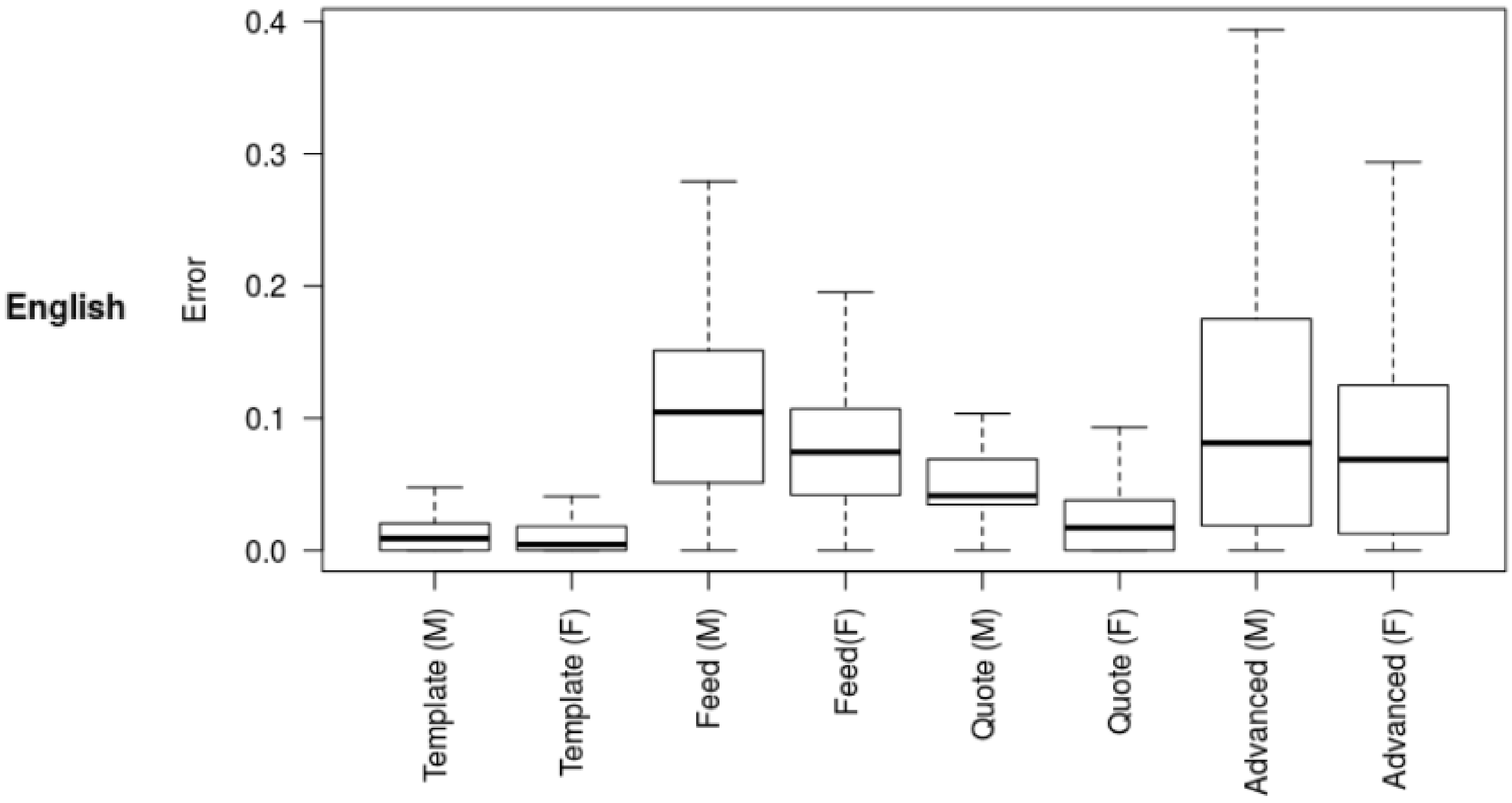
SPANISH



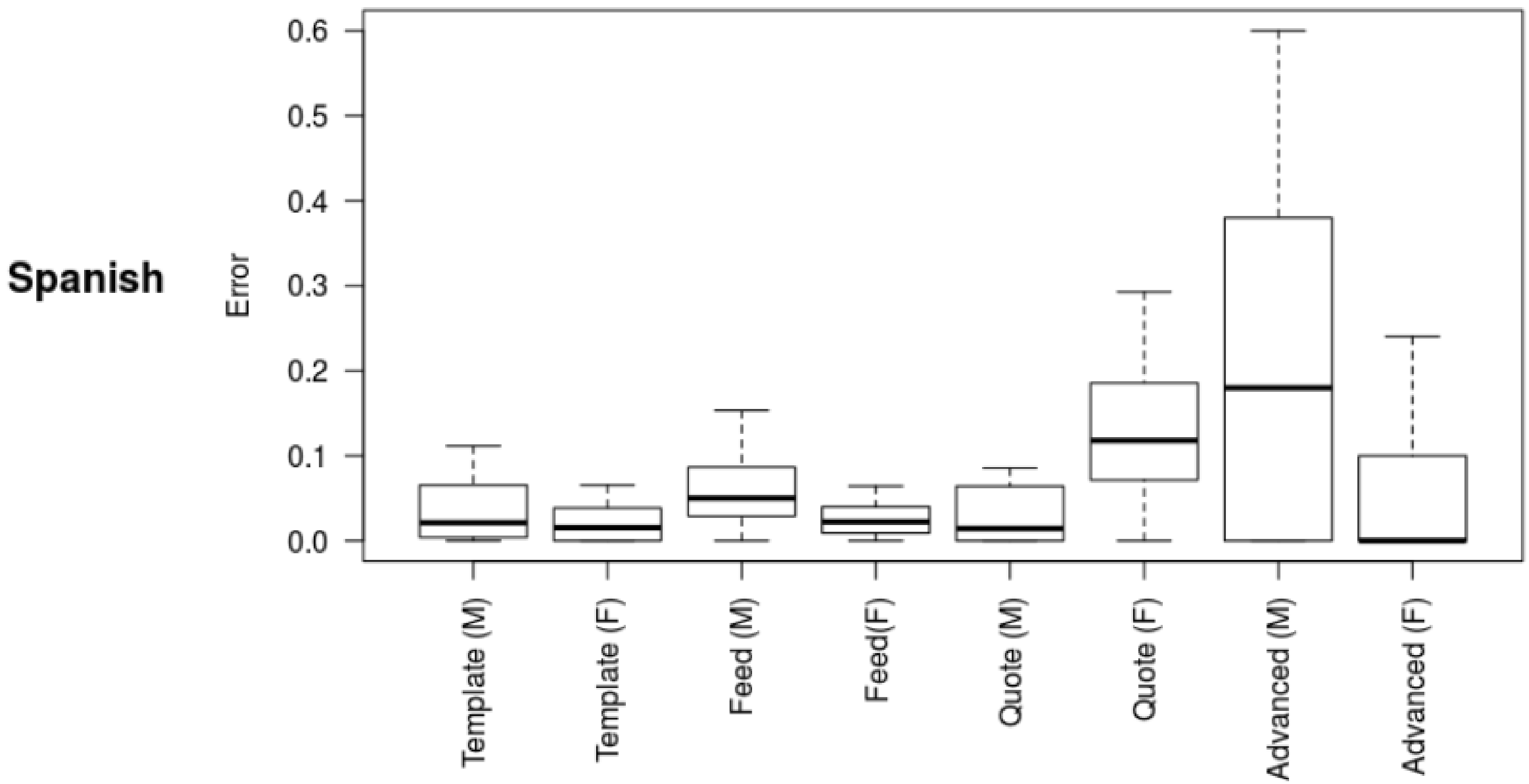
Errors per bot type

Author Id.	Twitter Account	Type	N. Systems
caf6d82d5dca1598beb5bfac0aea4161	@NasaTimeMachine	template	21 / 53
@wylejw You must be cool, I'll follow you!			
4c27d3c7a10964f574849b6be1df872d	@rarehero	feed	52 / 53
Get a doll, drape fabric and spray the hell out of it with Fabric Quick Stiffening Spray ... https://t.co/C9Ub6xXZWI via @duckduckgo			
8d08e3a0e1fea2f965fd7eb36f3b0b07	@MessiQuote	quote	48 / 53
.@PedroPintoUEFA: "Messi is unstoppable and we should feel privileged to be watching a player who may be the best of all time." https://t.co/TmCR6qCzO2			
6a6766790e1f5f67813afd7c0aa1e60d	@markov_chains	advanced	42 / 53
I have transferred to the local library go you! Just be Crazy John's prepaid sim card.			

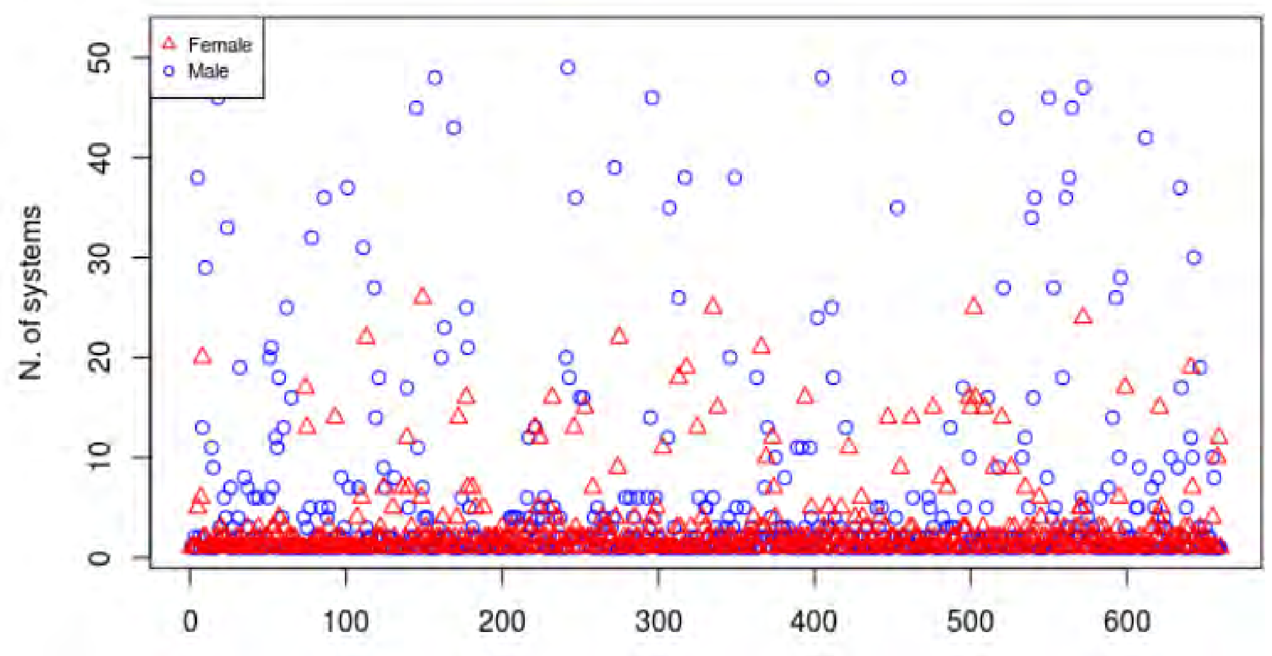
Bot to human per gender errors



Bot to human per gender errors



Human to bot errors



Human to bot errors

Author Id.	Twitter Account	Gender	N. Systems
63e4206bde634213b3a37343cf76e900	@Ask_KFitz	male	49 / 53
#Electric Imp Smart Refrigerator https://t.co/qigh5Womd7 https://t.co/JNVsRKvRQ8			
b11ffeeed0b38eb85e4e288f5c74f704	@iqbalmustansar	male	45 / 53
Trend - What's Dominating Digital Marketing Right Now? - https://t.co/dWp7ovqzCM			
ba0850ae38408f1db832707f1e0258fd	@CharBar_tweets	female	26 / 53
Hollywood boll #bowling #legs #Sundayfunday https://t.co/cLq9ZlNM38			
d64be10ecfbbb81d0c6e5b3115c335a5	@RheaRoryJames	female	25 / 53
RT @realDonaldTrump: Employment is up, Taxes are DOWN. Enjoy!			

https://botometer.iuni.iu.edu

Human to bot errors

Author Id.	Twitter Account	Gender	N. Systems
a22edd53bb04de0c06a52df897b13dd0	@carlosguadian	male	39 / 42
Tres días para analizar el presente y futuro de la Administración pública: lo que trae el Congreso NovaGob 2018 - NovaGob 2018 https://t.co/Ofc4cDTeym #novagob2018			
cf520c8e810a6a9bae9171d6f23c29be	@kicorangel	male	35 / 42
Google prepara una versión de pago para Youtube http://t.co/UvZdao68wc			
8e4340e95667c8add31f427a09dd3840	@EmaMArredondoM	female	30 / 42
@andrespino007 ¿Se ha preguntado cómo alguien llega a ser científico? Pequeña muestra chilena: https://t.co/fLjJsV0I0J			
6730bdf9686769c4a8a79d2f766a7f67	@Annie$_{H}go$	female	24 / 42
Wow!! Nuevamente rebasamos expectivas... https://t.co/PJ8bHA1SrG			

Conclusions

- Several approaches to tackle the task:
  - Best approach: n-grams + SVM
- Best results in **bots vs. human**:
  - **Over 84% on average** (EN 86.15%; ES 84.08%)
  - English (95.95%): Johansson - Stylistic features + Random Forest
  - Spanish (93.33%): Pizarro - n-grams + SVM
- Error analysis:
  - **Highest confusion from bots to humans (17.15% vs. 7.86% EN; 14.45% vs. 14.08% ES)**
    - ...**mainly towards males** (9.83% vs. 7.53% EN; 8.50% vs. 5.02% ES)
    - ...**males more confused with bots** (8.85% vs. 3.55% EN; 18.93% vs. 11.61% ES)
  - **Error per bot type**:
    - **Advanced bots: 30.11% EN; 32.38% ES**
    - EN: quote (12.64%); template (17.94%); feed (27.89%)
    - ES: quote (26.51%); template (13.20%); feed (14.28%)
    - **Mainly towards males**, except quote bots in ES (6.75% vs. 15.29% towards males)

Conclusions

Looking at the results, we can conclude:

- It is feasible to automatically identify bots in Twitter with high precision
  - ...even when **only textual features** are used.

- There are specific cases where the task is difficult due to:
  - ...**the language used by the bots (e.g., advanced bots)**
  - **...the way the humans use the platform (e.g., to share news)**

In both cases, although the precision is high, a major effort needs to be made to take into account **false positives**.

Industry @ author profiling

Organisation

autoritas
nuevas ideas, nuevas soluciones

Sponsors

THE LOGIC VALUE

IBM PartnerWorld

symanto GROUP

Industry @ author profiling

Organisation

autoritas
nuevas ideas, nuevas soluciones

Participants

FOI
Swedish Defence Research Agency

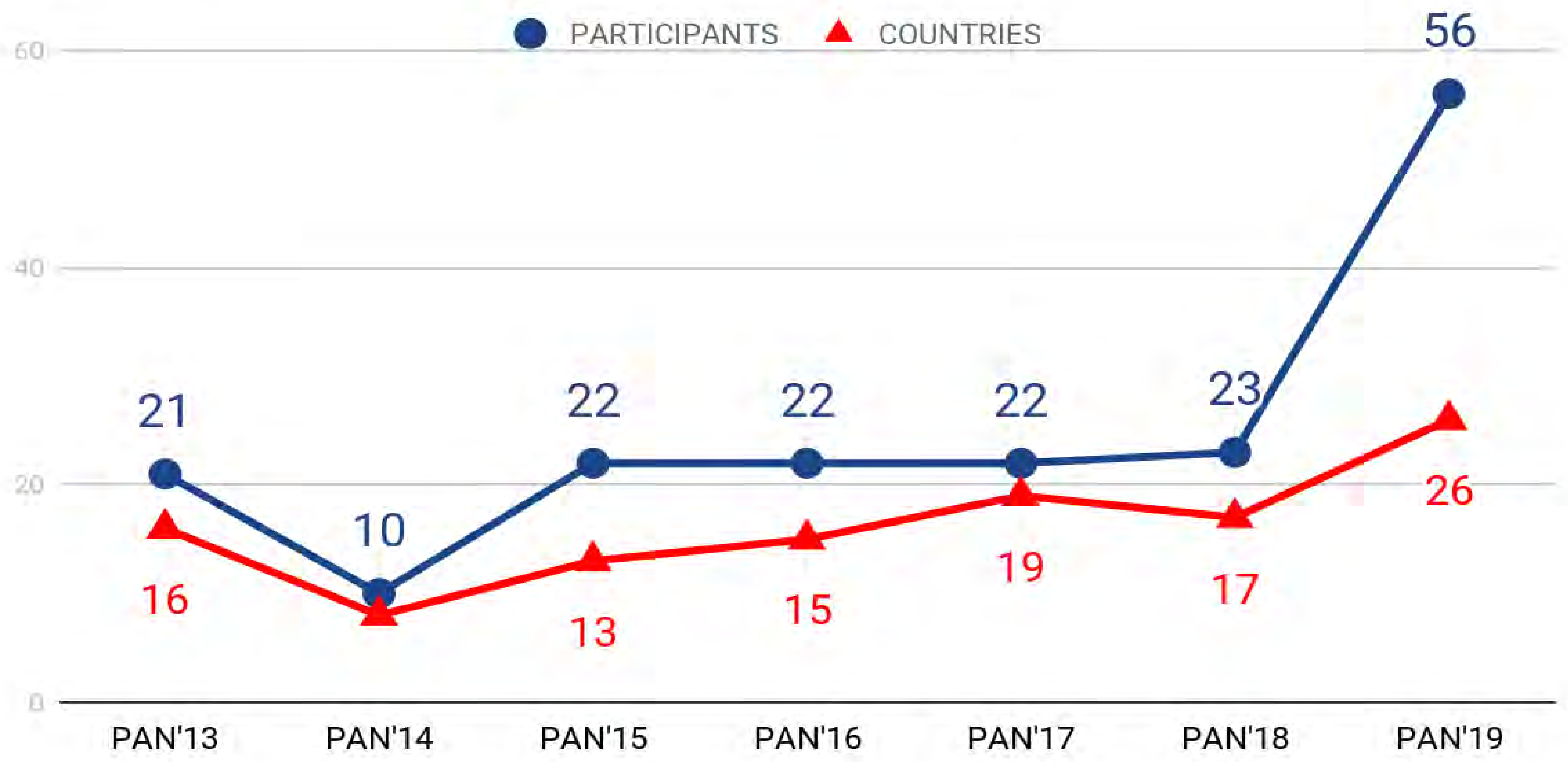
Sponsors

THE LOGIC VALUE

IBM PartnerWorld

symanto GROUP

Task impact


PARTICIPANTS
COUNTRIES
60
40
20
0
21
10
22
22
22
23
56
16
13
15
19
17
26
PAN'13
PAN'14
PAN'15
PAN'16
PAN'17
PAN'18
PAN'19

On behalf of the author profiling task organisers:

Thank you very much for participating
and hope to see you next year!!

Analysis of FAKE NEWS followers in Twitter