9th Author Profiling task at PAN

Profiling Hate Speech Spreaders on Twitter

PAN-AP-2021 CLEF 2021
Online, 21-24 September

Francisco Rangel - Symanto Research
Grétel Liz de la Peña Sarracén - Universitat Politècnica de València
BERTa Chulvi - Universitat Politècnica de València
Elisabetta Fersini - Università Degli Studi di Milano-Bicocca
Paolo Rosso - Universitat Politècnica de València

Introduction

Author profiling aims at identifying **personal traits** such as age, gender, personality traits, native language, language variety... from writings?

This is crucial for:

- Marketing.
- Security.
- Forensics.

Last years goal

Profiling Harmful Information Spreaders:

2019 - Profiling Bots
2020 - Profiling Fake News Spreaders
2021 - Profiling Hate Speech Spreaders

Task goal

Given a Twitter feed, determine whether its author is **keen to spread hate speech or not**.

Two languages:

English **Spanish**

Corpus

Methodology

1. Selection of users considered potential haters:
    1.1. Keyword-based search (hateful words mainly towards women or immigrants)
    1.2. User-based search (users appearing in reports and/or press) + following their networks
2. Timeline collection
    3.1. Manual review of the tweets conveying hate speech
    3.2. Users with more than ten hateful tweets are labelled as keen to spread them. Otherwise, they are not.

	(EN) English			(ES) Spanish		
	Keen to spread hate speech	Not keen to spread hate speech	Total	Keen to spread hate speech	Not keen to spread hate speech	Total
Training	100	100	200	100	100	200
Test	50	50	100	50	50	100
Total	150	150	300	150	150	300

For each user, we provided 200 tweets

Evaluation measures

The **accuracy** is calculated per language and averaged:

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

Baselines

RANDOM	A baseline that randomly generates the predictions among the different classes
CHAR N-GRAMS	With values for n from 2 to 6, with Logistic Regression
WORD N-GRAMS	With values for n from 1 to 3, with Support Vector Machines
Symanto (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: hate speech spreader / non-spreader. 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
USE	Universal Sentence Encoder to feed a BiLSTM
XLMR	XLM-Roberta transformer to feed a BiLSTM
MBERT	Multilingual BERT transformer to feed a BiLSTM
TFIDF	TFIDF vectors representing each user's text to feed a BiLSTM

Participation

66+1 participants
32 working notes
16+1 countries

https://mapchart.net/world.html

Approaches

What kind of ...

Preprocessing

Features

Methods

... did the teams perform?

Approaches - Preprocessing

Twitter elements (RT, VIA, FAV)	Del Campo et al.; Huertas-García et al.; Schlicht & de Paula; Jain et al.; Giglou et al.; Vogel & Meghana
Emojis and other non-alphanumeric chars	Alcañiz & Andrés; Jain et al.; Anwar; Giglou et al.; Vogel & Meghana; Bagdon; Espinosa & Sidorov; Cabrera et al.; Das & Patra
Lemmatisation	Vogel & Meghana; Alcañiz & Andrés; Das & Patra; Jain et al.
Tokenisation	Das & Patra; Jain et al.
Punctuation signs	Cabrera et al.; Puertas & Martínez-Santos; Dukic & Krzic; Espinosa & Sidorov; Giglou et al.; Alcañiz & Andrés; Del Campo et al.; Das & Patra; Jain et al.
Numbers	Del Campo et al.; Huertas-García et al.; Giglou et al.; Vogel & Meghana
Lowercase	Alcañiz & Andrés; Del Campo et al.; Das & Patra; Anwar; Vogel & Meghana; Bagdon; Dukic & Krzic
Stopwords	Alcañiz & Andrés; Das & Patra; Jain et al.; Vogel & Meghana
Character flooding	Bagdon; Vogel & Meghana; Del Campo et al.; Alcañiz & Andrés
Short texts	Vogel & Meghana

Approaches - Preprocessing

Contractions expansion	Alcañiz & Andrés
Colloquial tokens (slang)	Alcañiz & Andrés; Huertas-García et al.
t-SNE	Finogeev et al.; Ceron & Casula
Kolmogorov-Smirnov test	Ceron & Casula
TFIDF	Ikae
Shift graphs	Ikae

Approaches - Features

Stylistic features: - Number of occurrences - Verbs, adjs, pronouns - Number of hashtags, mentions, URLs... - Capital vs. lower letters - Punctuation marks - ...	Katona et al.; Zaragozá & Pinto
N-gram models	Andrade & Gonçalves; Das & Patra; Siino et al.; Ikae; Balouchzahi; Espinosa & Sidorov; Alcañiz & Andrés; Katona et al.
Emotional and personality features	Katona et al.; Cervero; Puertas & Martínez-Santos; Huertas-García et al.
Specialised lexicons (HS)	Lai et al.; Tosev & Gievska
Embeddings	Puertas & Martínez-Santos; Zaragozá & Pinto; Alcañiz & Andrés; Del Campo et al.
...Lexical+statistical+syntactical+phonetical	Puertas & Martínez-Santos
...Semantic-emotion-based	Cabrera et al.

Approaches - Features

Transformers	
...BERT	Huertas-García et al.; Finogeev & Kaprielova; Akomeah et al.; Alzahrani & Jololian; Anwar; Ceron & Casula; Uzan & Hacohen-Kerner; Dukic & Krzic
...SBERT	Schlicht & de Paula; Vogel & Meghana
...ALBERT	Akomeah et al.
...RoBERTa	Uzan & Hacohen-Kerner; Anwar; Bagdon
...BERTTweet	Anwar
...BETO	Anwar

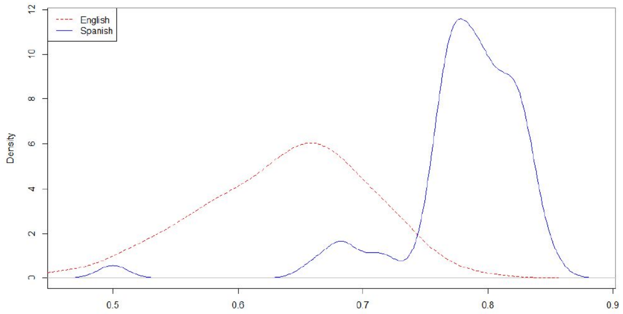
LDSE at char level	Babaei et al.
Fine-tuned transformer to modify the Impostor Method	Labadie et al.
Topics aggregation combined with ELMo to represent the user	Irani et al.
CNN to extract features from external data	Höllig et al.
Phonetic embeddings (among others)	Puertas & Martínez-Santos

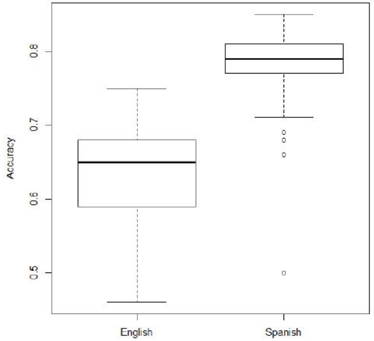
Approaches - Methods

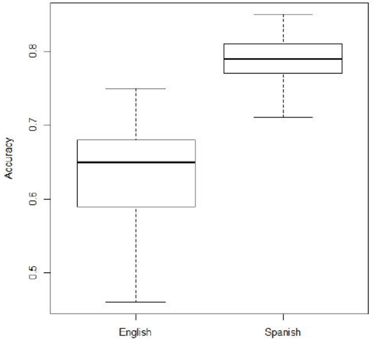
SVM	Alcañiz & Andrés; Del Campo et al.; Andrade & Gonçalves; Finogeev & Kaprielova; Höllig et al.; Das & Patra; Giglou et al.; Ceron & Casula; Vogel & Meghana; Badgon; Espinosa & Sidorov; Cabrera et al.
Logistic regression	Finogeev & Kaprielova; Badgon; Dukic & Krzic
Random Forest	Irani et al.; Puertas & Martínez-Santos; Andrade & Gonçalves
Ensembles	Huertas-García et al.; Cervero; Ikae; Balouchzahi et al.; Tosev & Gievska
Adaboost, Ridge, Naive Bayes, KNN, XGBoost, AutoML, ...	Katona et al.; Anwar; Jain et al.; Höllig et al.; Puertas & Martínez-Santos
Custom architecture	Martin et al.; Baris & Magnossao
RNN	Pallares & Herrero
CNN	Siino et al.
LSTM	Uzan & HacoHen-Kerner
bi-LSTM	Vogel & Meghana

Global ranking

STAT	EN	ES	AVG
Min	0.4600	0.5000	0.4800
Q1	0.5925	0.7700	0.6800
Median	0.6500	0.7900	0.7150
Mean	0.6377	0.7798	0.7066
SDev	0.0643	0.0539	0.0524
Q3	0.6800	0.8100	0.7400
Max	0.7500	0.8500	0.7900
Skewness	-0.4719	-2.6820	-1.4672
Kurtosis	2.7391	13.4989	7.1170
Normality (p-value)	0.2264	1.129e-08	0.0148







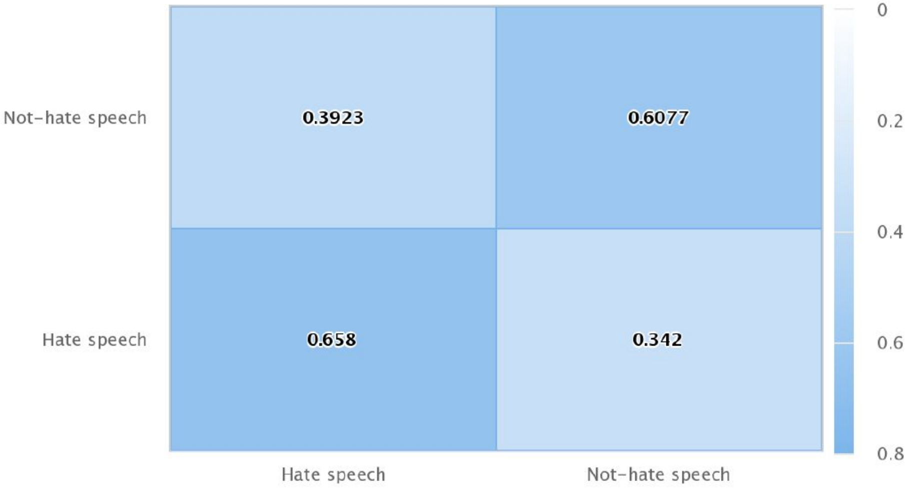
	PARTICIPANT	EN	ES	AVG
1	SiinoDiNuovo [66]	0.730	**0.850**	**0.790**
2	MUCIC [67]	0.730	0.830	0.780
2	UO-UPV [78]	0.740	0.820	0.780
4	andujar	0.720	0.820	0.770
4	anitei	0.720	0.820	0.770
4	anwar [54]	0.720	0.820	0.770
7	pagnan [62]	0.730	0.800	0.765
	LDSE [20]	0.700	0.820	0.760
	char nGrams+LR	0.690	0.830	0.760
8	hoellig [80]	0.730	0.790	0.760
9	bañuls	0.680	0.830	0.755
9	supaca	0.690	0.820	0.755
9	oleg [61]	0.670	0.830	0.750
9	moreno [53]	0.690	0.810	0.750
9	cervero [69]	0.700	0.800	0.750
14	katona [68]	0.700	0.790	0.745
	word nGrams+SVM	0.650	0.830	0.740
15	bagdon [55]	0.670	0.810	0.740
15	das [58]	0.670	0.810	0.740
17	ikae [63]	0.660	0.810	0.735
17	mata	0.700	0.770	0.735
19	lai [72]	0.620	0.840	0.730
19	jain [50]	0.660	0.800	0.730
19	villarroya	0.670	0.790	0.730
19	mktung	0.640	0.820	0.730
19	sercopa	0.670	0.790	0.730
19	castro	0.670	0.790	0.730
25	giglou [51]	0.650	0.800	0.725
25	huertas [48]	0.670	0.780	0.725
25	wentao	0.680	0.770	0.725
28	rus	0.610	0.830	0.720
28	tudo	0.650	0.790	0.720
30	jaiferhu	0.610	0.820	0.715
30	joshi	0.650	0.780	0.715
32	valiense [65]	0.630	0.790	0.710
32	krstev	0.650	0.770	0.710
34	martin [47]	0.650	0.770	0.710
35	gomez [74]	0.580	0.830	0.705
35	bakhteev	0.580	0.830	0.705

	Participant	En	Es	Avg
35	MaNa	0.640	0.770	0.705
38	cabrera [57]	0.620	0.780	0.700
38	esam [76]	0.630	0.770	0.700
38	zhang	0.630	0.770	0.700
41	dudko	0.610	0.780	0.695
41	meghana [52]	0.640	0.750	0.695
43	rubio	0.590	0.790	0.690
43	uzan [77]	0.620	0.760	0.690
45	herrero [81]	0.570	0.800	0.685
46	puertas [59]	0.600	0.760	0.680
	USE-LSTM	0.560	0.790	0.675
	XLMR-LSTM	0.620	0.730	0.675
47	ipek [49]	0.580	0.770	0.675
47	schlicht21	0.580	0.770	0.675
47	peirano	0.590	0.760	0.675
47	russo	0.550	0.800	0.675
	MBERT-LSTM	0.590	0.750	0.670
51	kazzaz	0.550	0.770	0.660
52	dorado	0.600	0.710	0.655
53	kobby [75]	0.530	0.770	0.650
53	kern	0.540	0.760	0.650
53	espinosa [56]	0.640	0.660	0.650
56	labadie	0.510	0.780	0.645
57	silva	0.560	0.690	0.625
57	garibo	0.570	0.680	0.625
59	estepicursor	0.510	0.720	0.615
60	spears	0.520	0.680	0.600
	TFIDF-LSTM	0.610	0.510	0.560
61	barbas	0.460	0.500	0.480

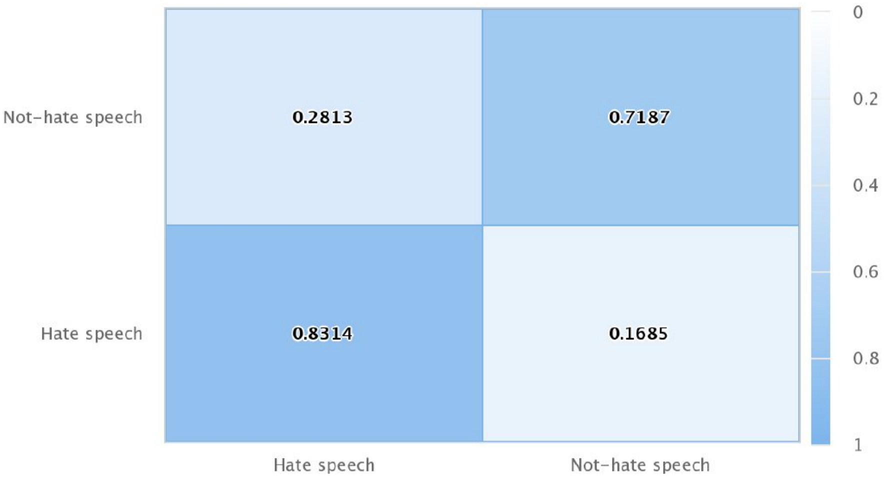
	Participant	En
62	dukic [60]	**0.750**
63	tosev [73]	0.700
64	amir [79]	0.680
65	siebert	0.680
66	iteam	0.650

Confusion matrices

ENGLISH



SPANISH



Best results at PAN'21

Dukic and Sovic
- BERT
- Logistic Regression

Siino et al.
- 100-dim word-embedding
- CNN

English		Spanish	
Dukić and Sović [60]	(0.75)	Siino et al. [66]	(0.85)

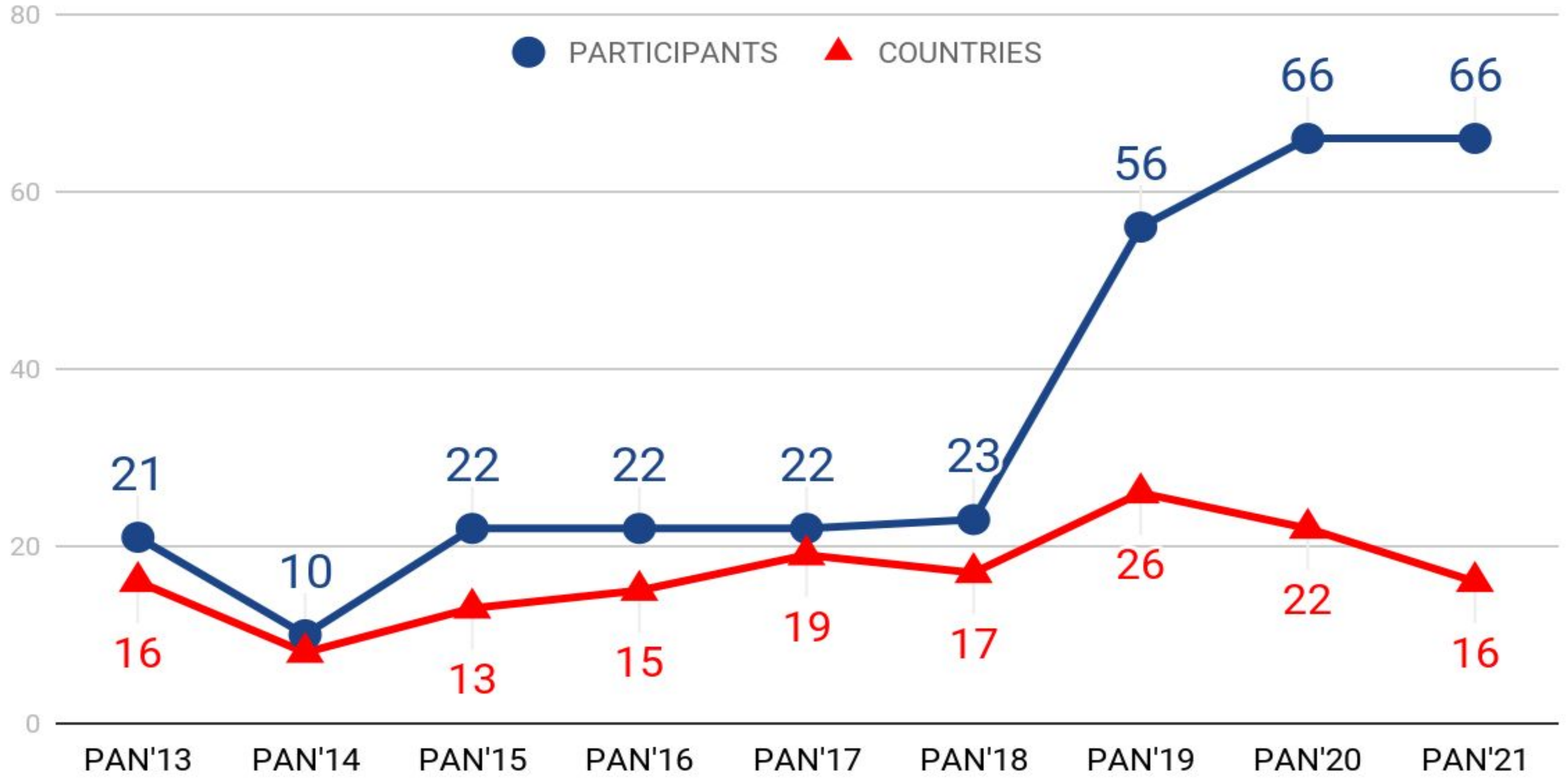
Conclusions

- Several approaches to tackle the task:
  - Traditional machine learning methods (SVM, LR) combined with BERT obtained the highest results.
- Results in English:
  - Over 64% on average.
  - Best (75%): Dukic and Sovic - BERT + Logistic Regression
- Results in Spanish:
  - Over 78% on average.
  - Best (85%): Siino et al. - 100-dimension word embedding + CNN
- Error analysis:
  - English:
    - False positives (non-hate speech spreaders as spreaders): 39.23%
    - False negatives (hate speech spreaders as non-spreaders): 34.40%
  - Spanish:
    - False positives (non-hate speech spreaders as spreaders): 28.13%
    - False negatives (hate speech spreaders as non-spreaders): 16.85%

Looking at the results, we can conclude:

- It is feasible to automatically identify Hate Speech Spreaders with high precision
  - ...even when only textual features are used.
- We have to bear in mind false positives since, in English they sum up to forty percent of the total predictions and in Spanish they are almost double than false negatives, and misclassification might lead to ethical or legal implications.

Task Impact


80
60
40
20
0
PARTICIPANTS
COUNTRIES
21
10
22
22
22
23
56
66
66
16
13
15
19
17
26
22
16
PAN'13
PAN'14
PAN'15
PAN'16
PAN'17
PAN'18
PAN'19
PAN'20
PAN'21

Industry at PAN (Author Profiling)

Organisation

symanto
psychology ai

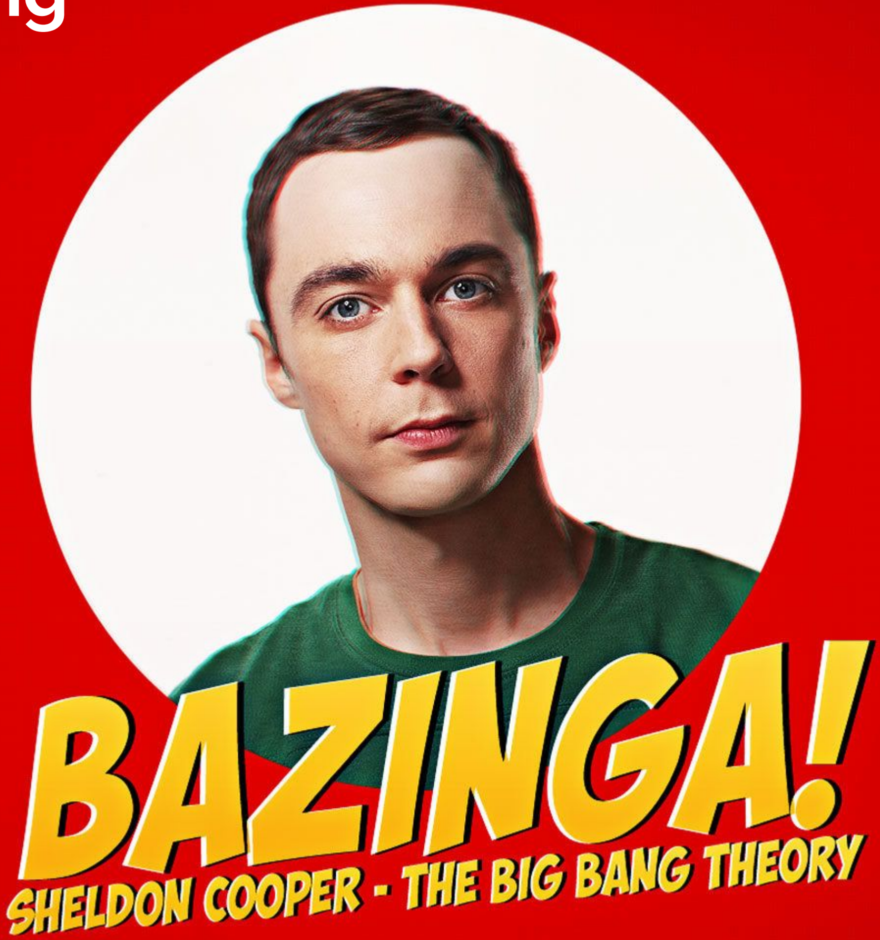
Sponsors

symanto
psychology ai

This year, the winners of the task are:

- Marco Siino, Elisa Di Nuovo, Ilenia Tinnirello and Marco La Cascia, Università degli Studi di Palermo and Università degli Studi di Torino, Italy

2022 -> Profiling Irony and Stereotypes spreaders



On behalf of the author profiling task organisers:

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