10th Author Profiling task at PAN

Profiling Irony and Stereotype Spreaders on Twitter

PAN-AP-2022 CLEF 2022
Bologna, 5-8 September

Reynier Ortega Bueno - Universitat Politècnica de València
Berta Chulvi - Universitat Politècnica de València
Francisco Rangel - Symanto Research

Paolo Rosso - Universitat Politècnica de València
Elisabetta Fersini - Università Degli Studi di Milano-Bicocca

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
2022 - Profiling Irony and Stereotypes Spreaders

Task goal

Given a Twitter feed, determine whether its author is **keen to spread irony or not,** with special focus on users who use irony towards **stereotypes**.

Language:

English

Corpus

Methodology

1. Defining Taxonomy and Stereotypes Categories:
   a. We examine the "target minority" field of the SBIC [Sap et al., 2020]
   b. We identify 600 unique labels that could be considered a social category in SBIC, and classify them into 17 categories:
   (1) national majority groups; (2) illness/health groups; (3) age and role family groups; (4) victims; (5) political groups; (6) ethnic/racial minorities; (7) immigration/national minorities; (8) professional and class groups; (9) sexual orientation groups; (10) women; (11) physical appearance groups; (12) religious groups; (13) style of life groups; (14) non-normative behaviour groups; (15) man/male groups; (16) minorities expressed in generic terms; (17) white people

2. Tweet Retrieval
   a. We retrieve tweets containing at least one of the labels included in **categories 5 to 14** of the taxonomy (with and without the hashtags irony and sarcasm).
   b. We select users with more tweets accomplishing previous conditions.

Corpus

Methodology

3. Annotation Process
   a. **Irony**: annotators were asked to mark the tweets where the user "express the opposite of what was saying as a disguised mockery". If the user had **more than 5 ironic tweets**, it was labelled as ironic.
   b. **Stereotypes**: annotators were asked to check if the social categories in the tweets were used to refer to a social group by associating them with a homogenising image of the category (e.g., as if all gays or Muslims were the same and could be described with that word). If the user had **more than 5 tweets with stereotypes**, it was labelled accordingly.

4. Corpus Construction
   a. Two independent annotators labelled the data (IAA **0.7093**).
   b. A third annotator sorted out disagreements.
   c. For each user, we provide 200 tweets.

Corpus

Statistics

	IRONIC			NON-IRONIC			Total
	Stereotypes	Non-stereotypes	Total	Stereotypes	Non-stereotypes	Total	
Training	140	70	210	140	70	210	420
Test	60	30	90	60	30	90	180
Total	200	100	300	200	100	300	600

Number of users per class and set. Each user contains 200 tweets.

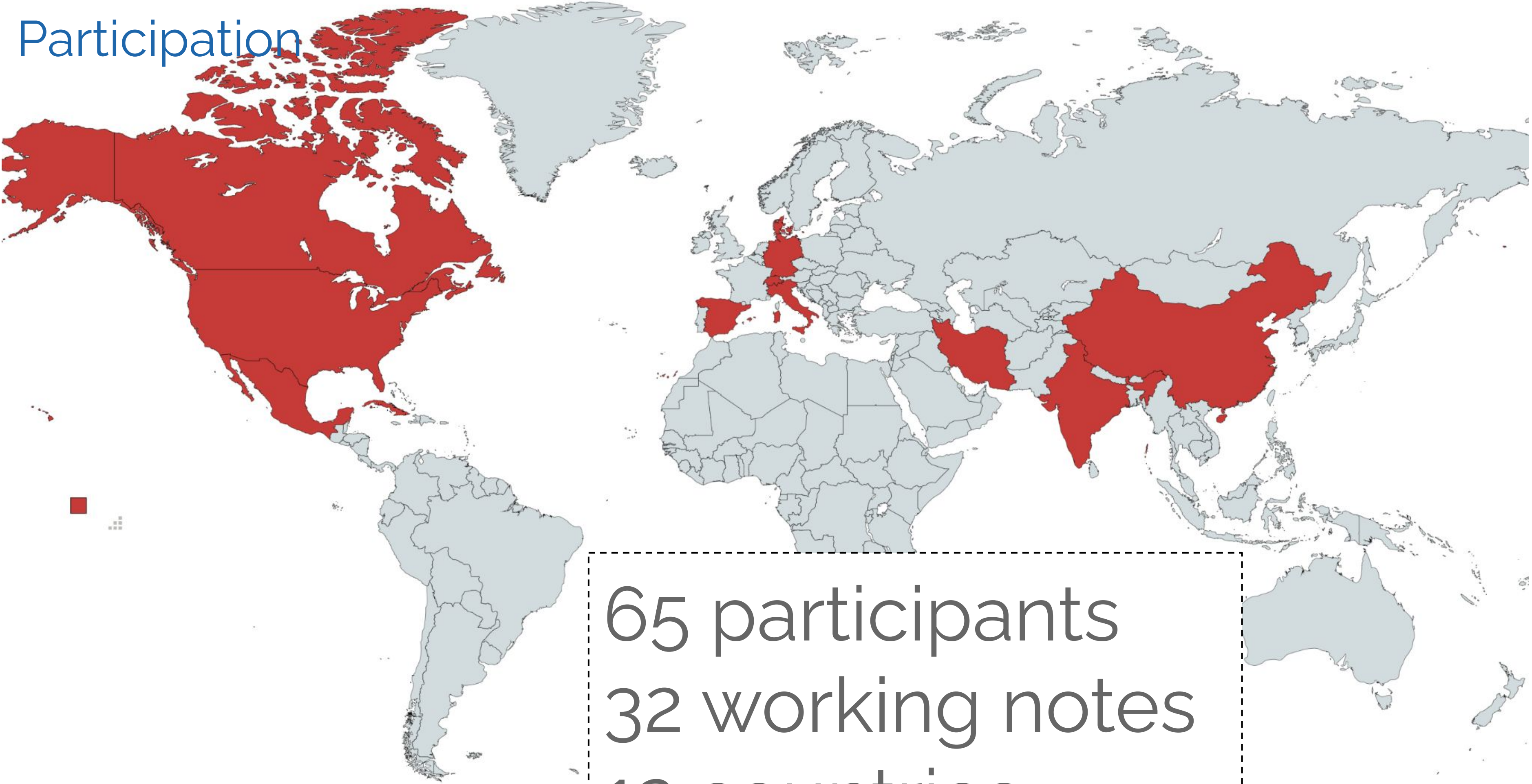
Evaluation measures

Since the dataset is completely balanced for the two target classes, ironic vs. non-ironic, we have used the **accuracy** measure and ranked the performance of the systems by that metric.

Baselines

CHAR 2-GRAMS + RF	Bigrams with Random Forest
WORD 1-GRAMS + LR	Bag of words with Logistic Regression
BERT + LSTM	We represent each tweet in the profile utilising pretrained Bert-base model. Later, we fed an LSTM with these vectors as input.
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: irony 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

Participation



65 participants
32 working notes
12 countries

https://mapchart.net/world.html

Approaches

What kind of ...

Preprocessing

Features

Methods

... did the teams perform?

Approaches - Preprocessing

Twitter elements (RT, VIA, FAV)	Giglou et al.; Cao et al.; Lin et al.; Jang; Xu et al.; Wang & Ning; Zhang & NIng; Sagar & Varma; Hazrati et al.
Emojis and other non-alphanumeric chars	Wang & Ning; Zhang & Ning; Hazrati et al.; Butt et al.; Jian & Huang
Lemmatisation	Haolong & Sun
Punctuation signs	Giglou et al.; Lin et al.; Xu & Ning; Wang & Ning; Hazrati et al.; Dong et al.
Numbers	Xu & Ning; Hazrati et al.; Butt et al.
Lowercase	Giglou et al.; Xu & Ning; Butt et al.; Dong et al.; Siino & Cascia; Mangione & Garbo; Zengyao & Zhongyuan; Croce et al.; Herold & Castro
Stopwords	Giglou et al.; Hazrati et al.; Butt et al.
Infrequent terms	Giglou et al.; Wang & Ning; Zengyao & Zhongyuan
X^2 with PMI and TF/IDF	Ikae
I/NI labels to the end of the tweets	Jian & Huang
GloVe to filter out features	Haolong & Sun

Approaches - Features

Stylistic features: - Vocabulary size - Number of tokens - Tweet length - Number of hashtags, mentions, URLS - Number of emojis - ...	Nikolova et al.
N-gram models	Sagar & Varma; Butt et al.; Herold & Castro; Ikae; Nikolova et al.
Emotional and personality features	Butt et al.
TextVectorizer	Croce et al.
Embeddings	Dong et al.; Yang et al.
Transformers	
...BERT	Cao et al.; Lin et al.; Xu & Ning; 71; Hazrati et al.; Jian & Huang; Zengyao & Zhongyuan; Wentao & Kolossa; Rodriguez & Barroso
...SBERT	Tahaei et al.
...BERTTweet	Wang & Ning

Approaches - Features

Transformers + others	
...BERT & Twitter RoBERTa + LM HateXPlain	Mathew et al.
SBERT + emojis	Tahaei et al.
SBERT + psychometrics, emotions and irony	Tavan et al.
BERT + TF-IDF n-grams	Das et al.; Gómez & Parres
SBERT + graph-based & one-hot embeddings	Giglou et al.
+ Ironic-, contextual-, psychometric-related features fine-tuned with datasets annotated with sentiment/emotions from Kaggle	Tavan et al.
Sentence Transformers + n-grams + stylistic	Jang
BERT & RoBERTa + FastText + stylistic	Díaz et al.

Approaches - Features

CNN	Díaz et al.
CNN + TF-IDF 1-grams + BiGRU	Haolong & Sun
Sequence probabilities + n-grams + GPT2 & DistilGPT2	Huang
Irony identification at tweet level by combining: 1) structural features 2) sentiment words 3) fine-grained emotions by means of several lexicons	Hernández & Montes

Approaches - Methods

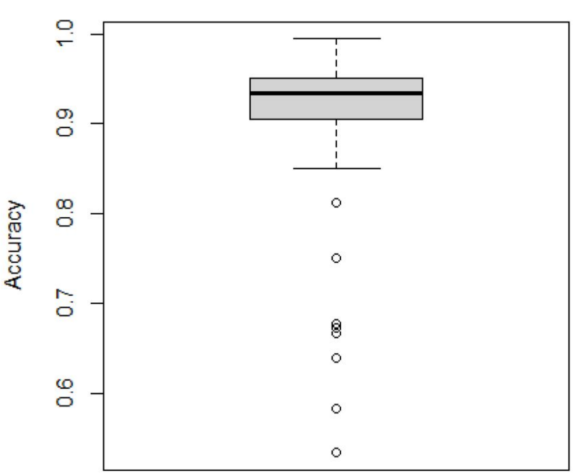
Logistic regression	Butt et al.; Das et al.
Random Forest	Sagar & Varma; Ikae; Nikolova et al.; Hernández & Montes
Bayes	Huang
Multilayer Perceptron	Hazrati et al.
Gradient Booster Classifier	Tavan et al.
k-Nearest Neighbours	Rodriguez & Barroso
Ensembles (trad. classifiers)	Zengyao & Zhongyuan; Herold & Castro; Tavan et al.
Ensembles (trad. Class + DL)	Siino & Cascia; Croce et al.

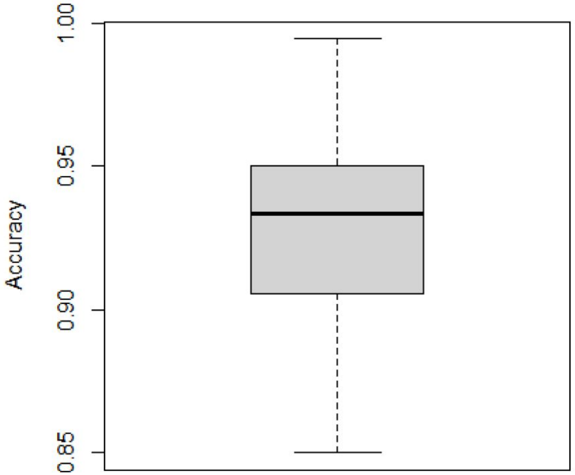
Approaches - Methods

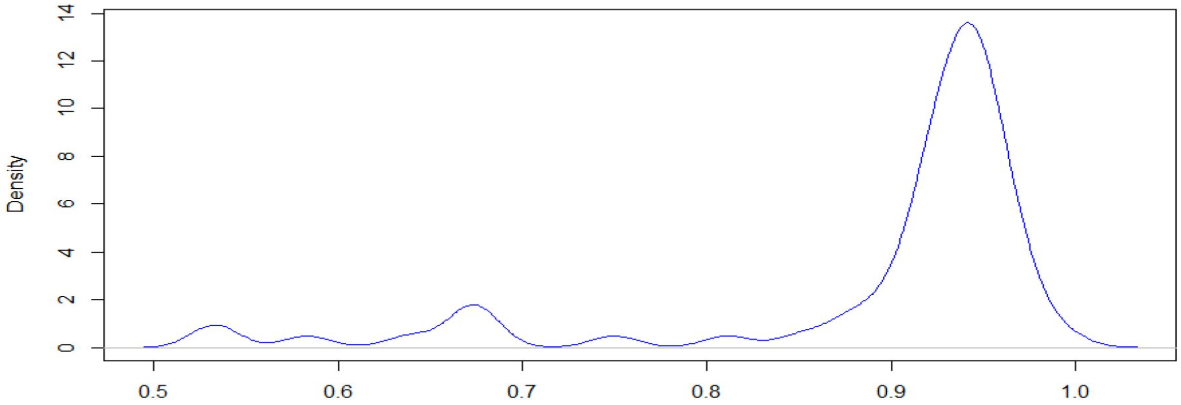
Linear Feed Forward Networks	Tahaei et al.
CNN	Dong et al.; Mangione & Garbo; Wentao & Kolossa
GCNN	Giglou et al.
bi-LSTM + CNN	Yang et al.
Fully-connected networks	Haolong & Sun; Labadie & Castro; Díaz et al.
AutoKeras; AutoGluon	Wang & Ning; Xu & Ning; Zhang & Ning
BERT + Decision Rules	Jian & Huang
BERT + SVM, MLP, Gaussian NB & RF	Rodriguez & Barroso
BERT + CNN, LSTM, att. layer	Gómez & Parres
BERT & DistilBERT + RF & SVM	Jang
BERT + voting classifier	Cao et al.; Lin et al.

Global ranking

Team	Acc
Best	0.9944
LDSE	0.9389
Char 2-grams + RF	0.8610
BoW + LR	0.8490
BERT + LSTM	0.6940
Worst	0.5333

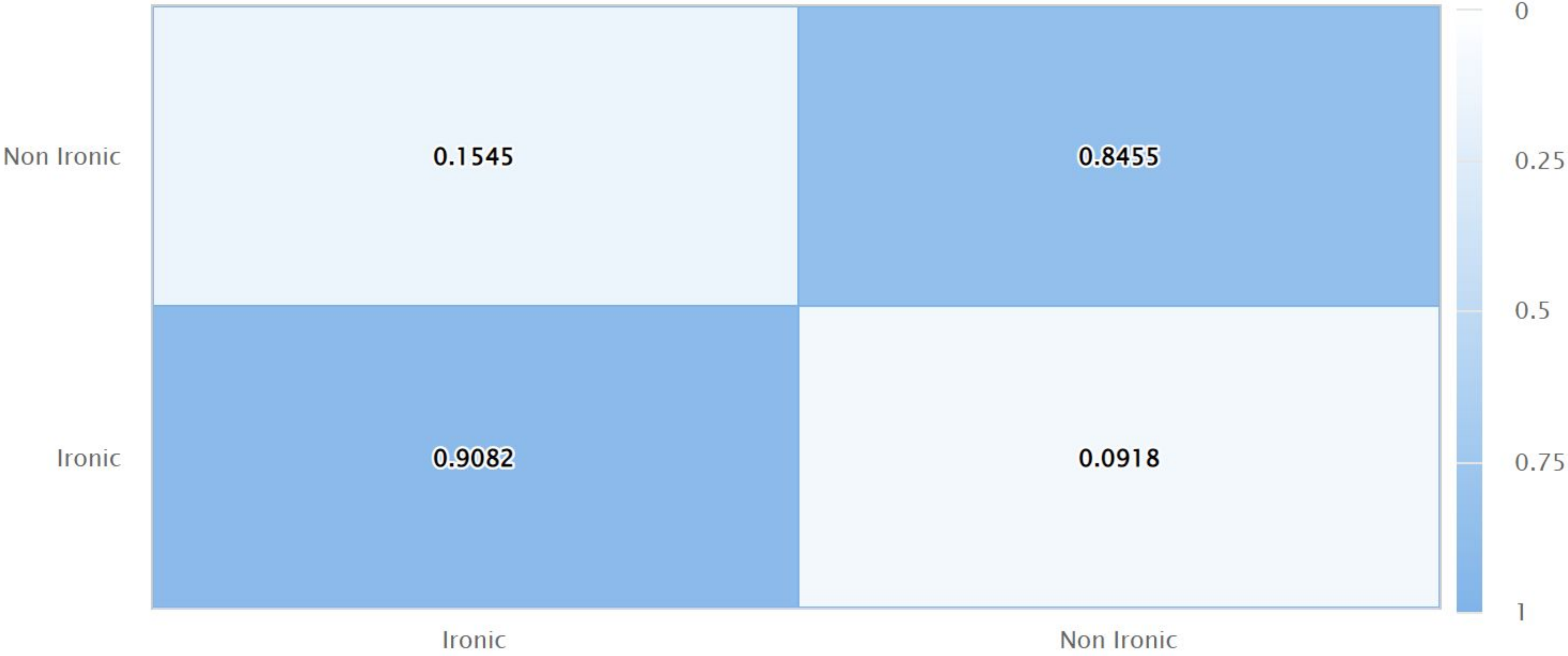






Min	Q1	Median	Mean	SDev	Q3	Max	Skewness	Kurtosis
0.5333	0.9056	0.9333	0.8926	0.1102	0.9500	0.9944	-2.0641	6.1113

Confusion matrix



	Ironic	Non Ironic
Non Ironic	0.1545	0.8455
Ironic	0.9082	0.0918

Corpus analysis

Does these high results mean that the corpus is biased?

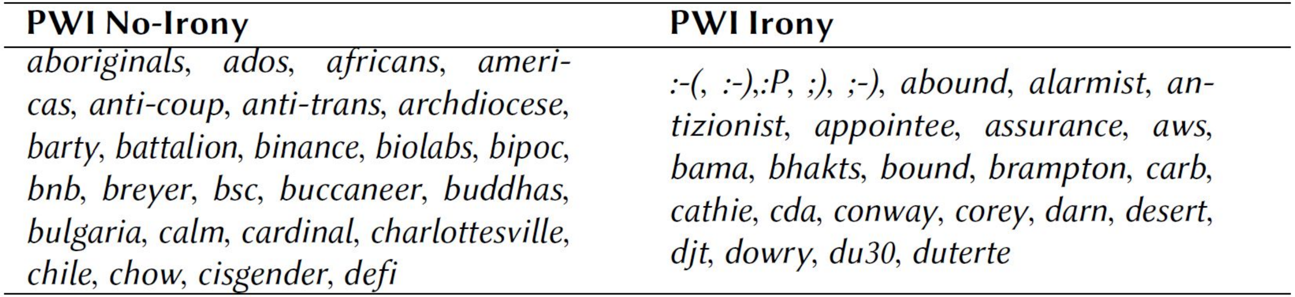
We have conducted a deep analysis of the corpus from five different angles:

- **Topics** used by ironic and non-ironic users.
- **Twitter elements** usage.
- **Language style**: categorical vs. narrative.
- **Emotionality**: activation, imaginary, pleasantness.
- **Personality** type and **communication style** of the users.

Topic-based analysis

Two-fold analysis:

- Determining the set of words that are highly polarized according to the indexes introduced in [Poletto et al., 2021]: Polarized Wiredness Index (PWI) which takes into account how polarized the words are in each class in the corpus (irony and non-irony).
- Determining the set of unique words in each class and analysing how this vocabulary impacts the learning process.



PWI No-Irony	PWI Irony
aboriginals, ados, africans, americas, anti-coup, anti-trans, archdiocese, barty, battalion, binance, biolabs, bipoc, bnb, breyer, bsc, buccaneer, buddhas, bulgaria, calm, cardinal, charlottesville, chile, chow, cisgender, defi	*:-(, :-),:P, ;), ;-), abound, alarmist, antizionist, appointee, assurance, aws, bama, bhakts, bound, brampton, carb, cathie, cda, conway, corey, darn, desert, djt, dowry, du30, duterte*

Related to ethnic

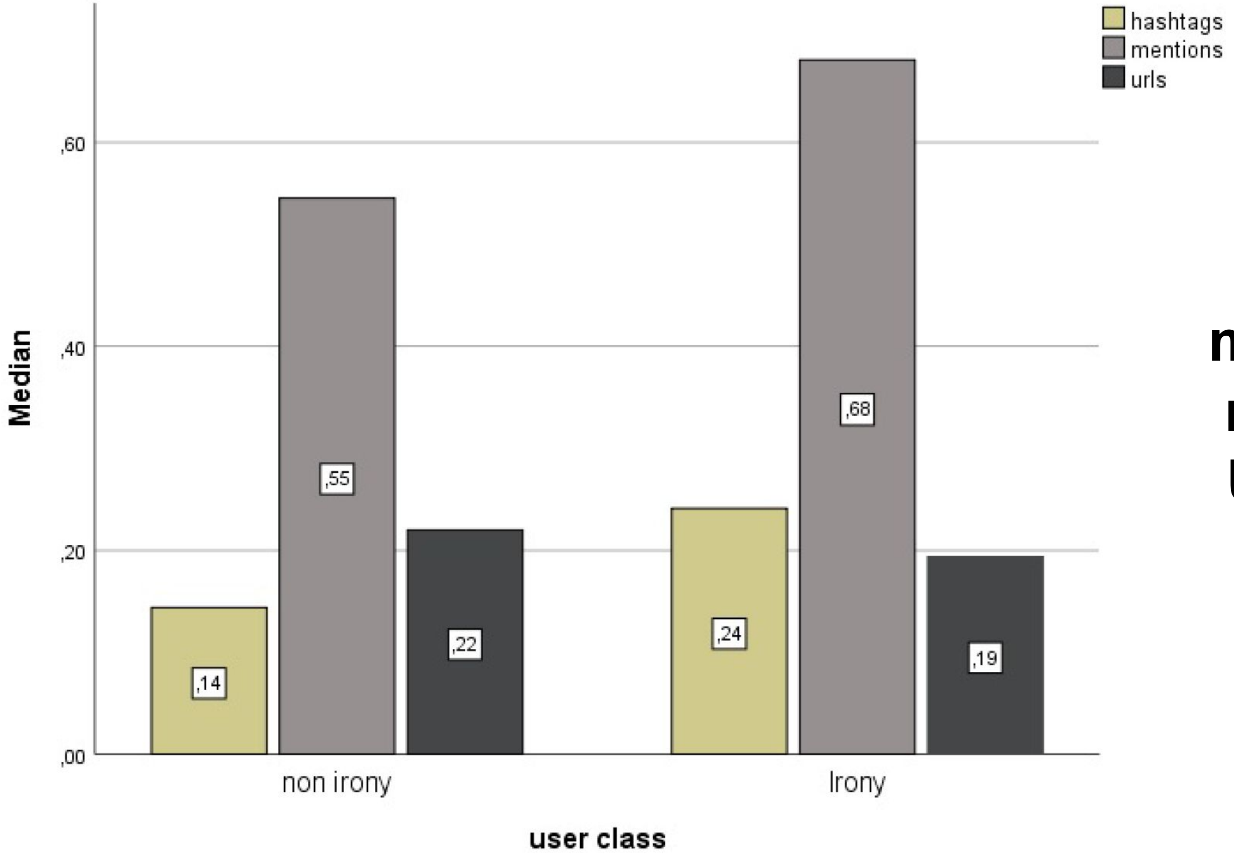
Related to politics and religion... and emojis

- We identified 1,379 words which only appear in one class.
- In the irony class, we found 334 unique terms, whereas, in the class non-irony, we identified 1,045 unique terms.

- We trained an SVM and RF classifiers considering as features the words in this vocabulary, obtaining respectively an accuracy of 0.8817 and 0.8763.

Highlights: there seems that there is a topic bias, but since there are also stylistic elements that vary between classes, this bias may be inherent in the type of users per class.

Twitter elements analysis



Ironic people write **shorter tweets**, use **more hashtags**, **more mentions**, and **fewer URLs** than non-ironic ones.

For all of them, the Mann-Whitney test is significant in both sets (training and test); p<.001

Language style analysis

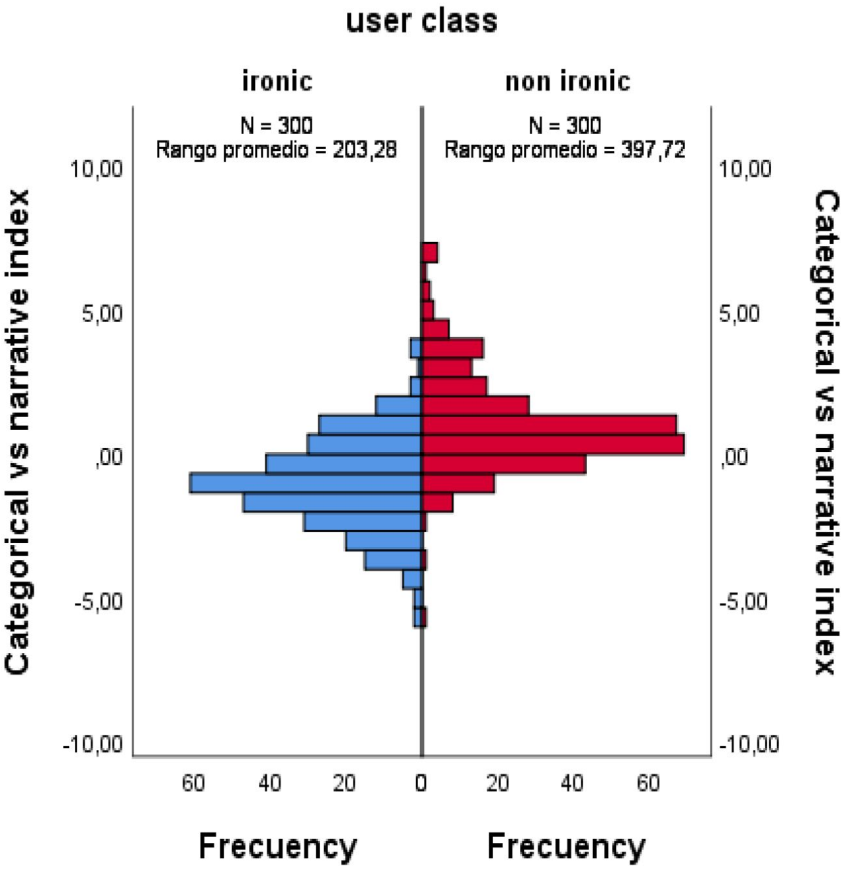
POS tagging with Freeling, and calculation of the Categorical (+) vs. Narrative (-) style inspired in [Nisbett et al., 2001] and composed as:

NOUN + ADJ + PREP - VERB - ADV - PERSPRON

- Categorical style is used to express ideas and concepts, whereas narrative is used to tell stories.

Highlights:

- Non-ironic users use significantly more a categorical style, while ironic users utilise more a narrative style (Mdn=-0.99; U=15,834; p<.001).
- Language style is topic agnostic: regardless of the topic, there are significant differences in the way ironic and non-ironic users employ language.



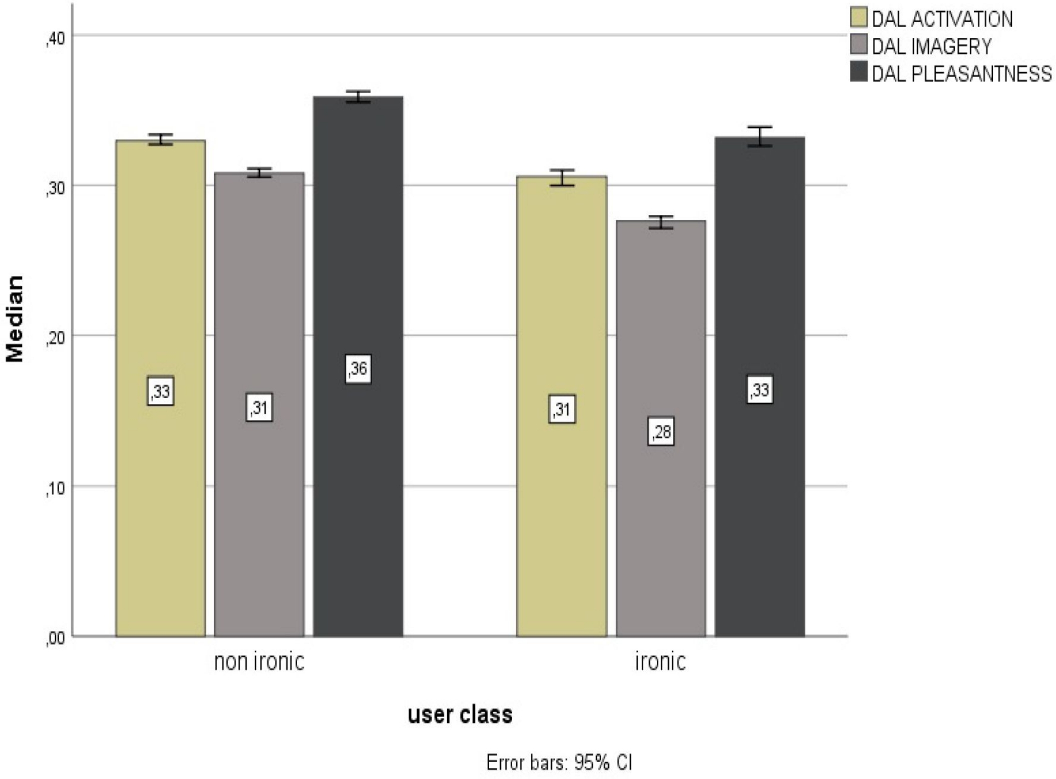
Emotions analysis

The new Dictionary of Affect in Language (DAL) [Whissell, 1989]:

- List of 8,742 words.
- Annotated by 200 naïve volunteers.
- Three dimensions: activation (active vs. passive), imaginary (easy vs. difficult to imagine), and pleasantness (unpleasant vs. pleasant).

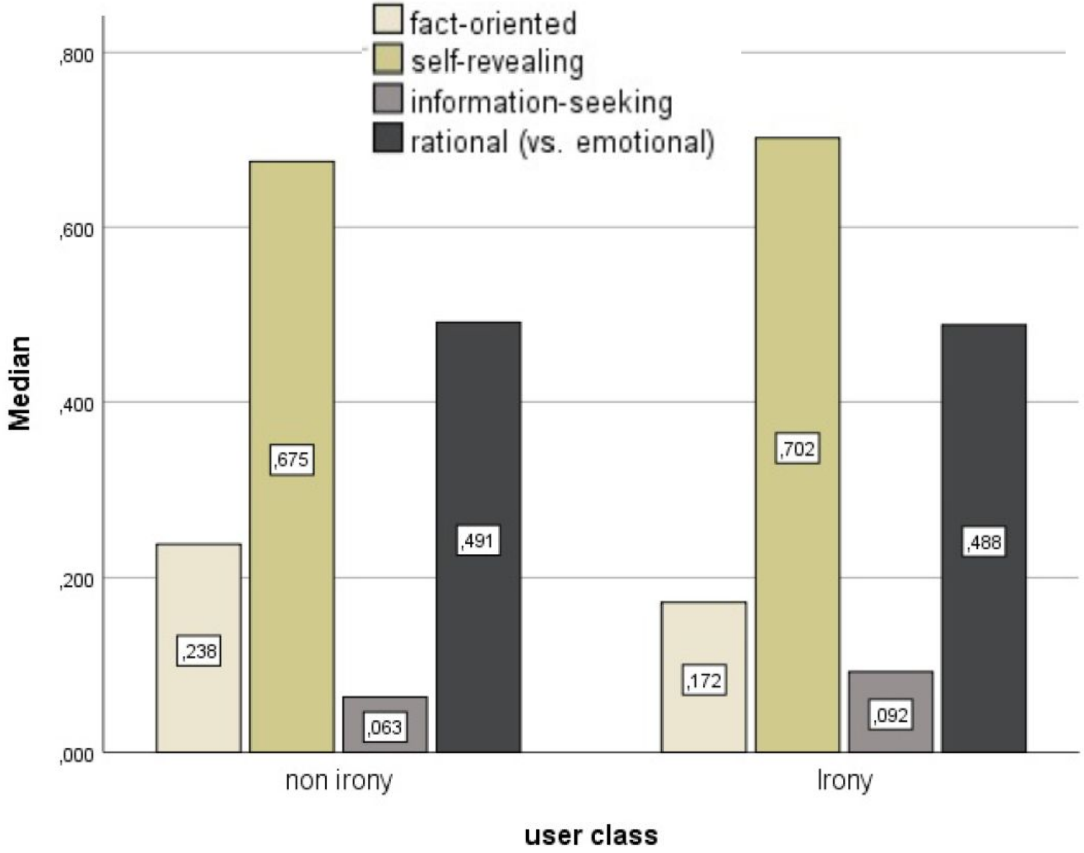
Highlights:

- Non-ironic users present higher scores in the three emotional dimensions than ironic ones.



For all of them, the Mann-Whitney test is significant in both sets (training and test); p<.001

Communication styles



For all of them, the Mann-Whitney test is significant p<.001

Symanto API* to obtain the users' personality type and communication styles [Stajner et al., 101]:

- The personality type refers to the way the person behaves in a specific interaction from the emotional vs rational point of view.
- Action-seeking, defined as direct or indirect requests, suggestions, and recommendations that expect action from other people.
- Fact-oriented, where the user utilises factual and objective statements.
- Self-revealing, when the users share personal information or experiences.
- Information-seeking, defined as direct or indirect questions searching for information.

Highlights:

- Ironic and non-ironic users do not differ in action-seeking but the do in the other four styles.
- Ironic users use less the fact-oriented style and more a self-revealing and information-seeking style.
- Non-ironic users are more emotional and less rational than ironic ones.

*https://rapidapi.com/collection/symanto-symanto-default-apis

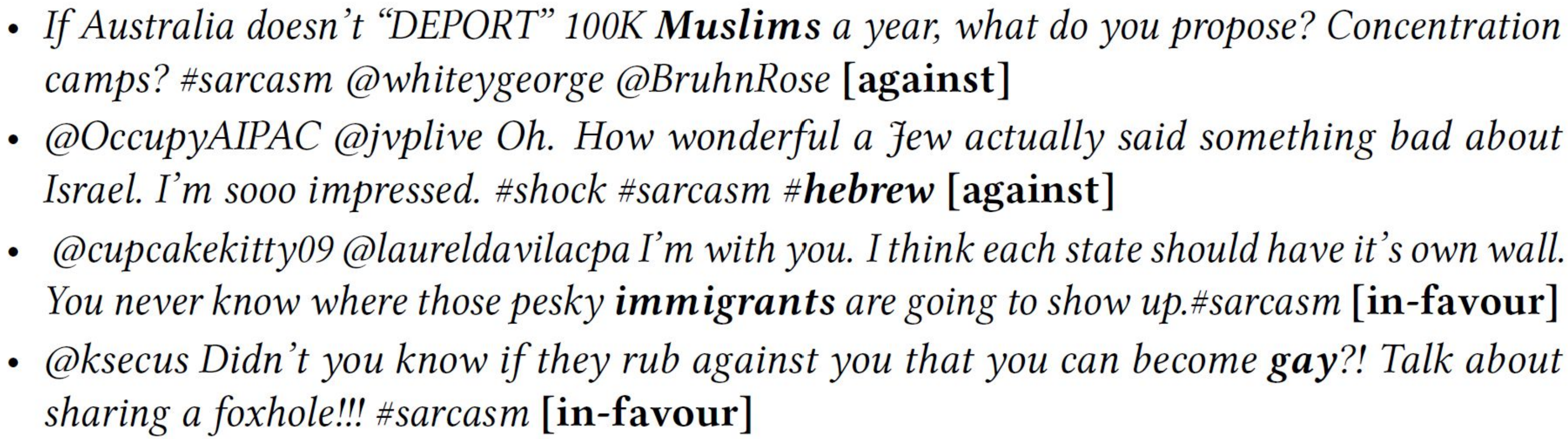
Subtask

The aim of the **Profiling Stereotype stance on Ironic Authors** subtask is to detect whether ironic users employed stereotypes to hurt the target or to somehow defend it.

Language:

English

Examples



- *If Australia doesn't "DEPORT" 100K* ***Muslims*** *a year, what do you propose? Concentration camps? #sarcasm @whiteygeorge @BruhnRose* **[against]**
- *@OccupyAIPAC @jvplive Oh. How wonderful a Jew actually said something bad about Israel. I'm sooo impressed. #shock #sarcasm #****hebrew*** **[against]**
- *@cupcakekitty09 @laureldavilacpa I'm with you. I think each state should have it's own wall. You never know where those pesky* ***immigrants*** *are going to show up.#sarcasm* **[in-favour]**
- *@ksecus Didn't you know if they rub against you that you can become* ***gay****?! Talk about sharing a foxhole!!! #sarcasm* **[in-favour]**

Corpus

Methodology

1. We selected those authors that were annotated as ironic and spreaders of stereotypes.
2. For each author, only the tweets marked as ironic and using stereotypes were annotated with their stance.
3. We asked the annotators to rely on their own perspectives on whether the tweets are in favour or against the mentioned social category, with no other guidance.
4. The overall stance of an author corresponds to the majority class at tweet level.
5. The IAA between the first two annotators was 0.645.
6. A third annotator sorted out disagreements.

	IN FAVOUR	AGAINST	Total
Training	46	94	140
Test	12	48	60
Total	58	142	200

For each user, we provided 200 tweets

Evaluation measures

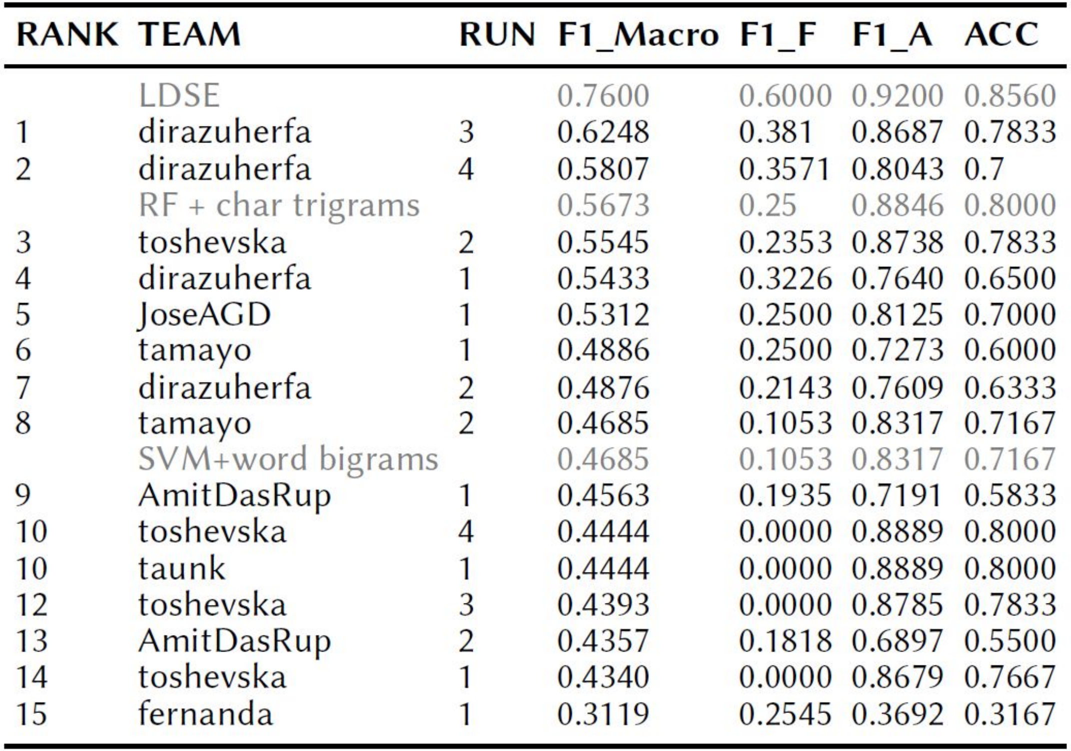
The performance of the systems is evaluated using the macro averaged F1 measure (F_Macro).

We also analyse the F1-measure per class to study more in depth the behaviour of the systems.

Baselines

CHAR 3-GRAMS + RF	Trigrams with Random Forest
WORD 2-GRAMS + SVM	Bigrams with Support Vector Machine
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: irony 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

Results

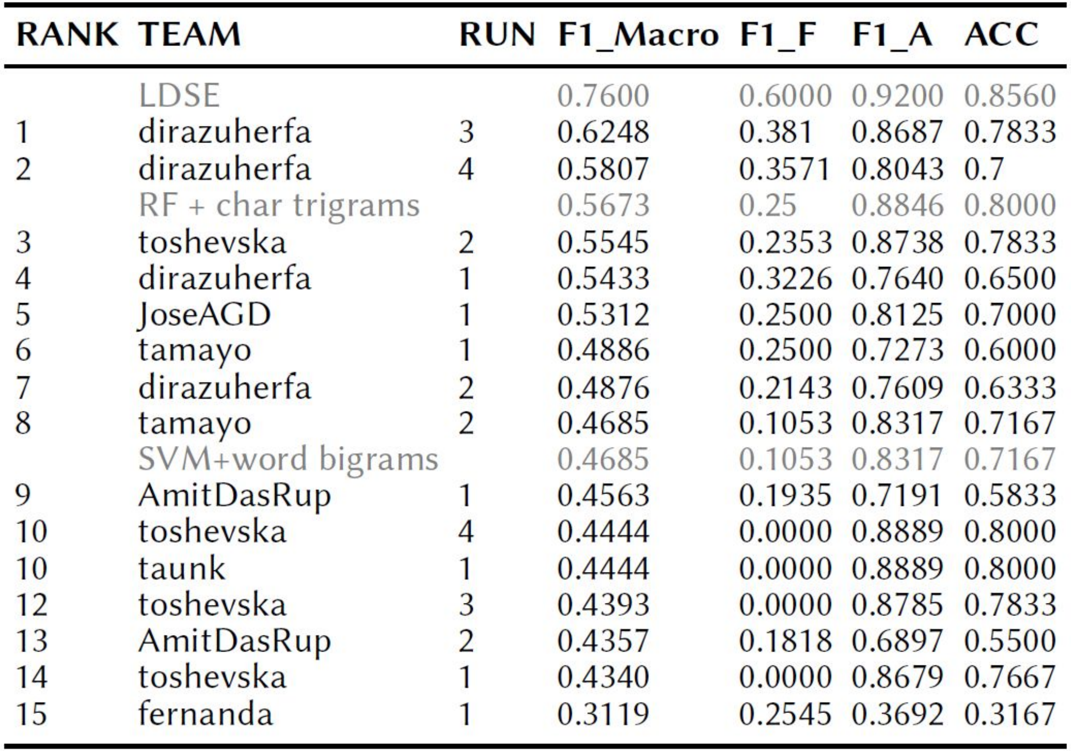


RANK	TEAM	RUN	F1_Macro	F1_F	F1_A	ACC
	LDSE		0.7600	0.6000	0.9200	0.8560
1	dirazuherfa	3	0.6248	0.381	0.8687	0.7833
2	dirazuherfa	4	0.5807	0.3571	0.8043	0.7
	RF + char trigrams		0.5673	0.25	0.8846	0.8000
3	toshevska	2	0.5545	0.2353	0.8738	0.7833
4	dirazuherfa	1	0.5433	0.3226	0.7640	0.6500
5	JoseAGD	1	0.5312	0.2500	0.8125	0.7000
6	tamayo	1	0.4886	0.2500	0.7273	0.6000
7	dirazuherfa	2	0.4876	0.2143	0.7609	0.6333
8	tamayo	2	0.4685	0.1053	0.8317	0.7167
	SVM+word bigrams		0.4685	0.1053	0.8317	0.7167
9	AmitDasRup	1	0.4563	0.1935	0.7191	0.5833
10	toshevska	4	0.4444	0.0000	0.8889	0.8000
10	taunk	1	0.4444	0.0000	0.8889	0.8000
12	toshevska	3	0.4393	0.0000	0.8785	0.7833
13	AmitDasRup	2	0.4357	0.1818	0.6897	0.5500
14	toshevska	1	0.4340	0.0000	0.8679	0.7667
15	fernanda	1	0.3119	0.2545	0.3692	0.3167

Methods:

- dirazuherfa: Emotion-based approach combining structural-, sentiment-, and emotion-based features.
- toshevska: Deep graph convolutional neural network.
- JoseAGD: UMUTextStats + FastText + BERT + RoBERTa & a fully-connected network.
- tamayo: RoBERTa + KNN to prototype creation.
- AmitDasRup: BERT + TFIDF & LR.
- taunk: TFIDF BoW + trad. ML methods.
- fernanda: n-grams combination + voting.

Results



RANK	TEAM	RUN	F1_Macro	F1_F	F1_A	ACC
	LDSE		0.7600	0.6000	0.9200	0.8560
1	dirazuherfa	3	0.6248	0.381	0.8687	0.7833
2	dirazuherfa	4	0.5807	0.3571	0.8043	0.7
	RF + char trigrams		0.5673	0.25	0.8846	0.8000
3	toshevska	2	0.5545	0.2353	0.8738	0.7833
4	dirazuherfa	1	0.5433	0.3226	0.7640	0.6500
5	JoseAGD	1	0.5312	0.2500	0.8125	0.7000
6	tamayo	1	0.4886	0.2500	0.7273	0.6000
7	dirazuherfa	2	0.4876	0.2143	0.7609	0.6333
8	tamayo	2	0.4685	0.1053	0.8317	0.7167
	SVM+word bigrams		0.4685	0.1053	0.8317	0.7167
9	AmitDasRup	1	0.4563	0.1935	0.7191	0.5833
10	toshevska	4	0.4444	0.0000	0.8889	0.8000
10	taunk	1	0.4444	0.0000	0.8889	0.8000
12	toshevska	3	0.4393	0.0000	0.8785	0.7833
13	AmitDasRup	2	0.4357	0.1818	0.6897	0.5500
14	toshevska	1	0.4340	0.0000	0.8679	0.7667
15	fernanda	1	0.3119	0.2545	0.3692	0.3167

Highlights:

- Low performance in the "in-favour" class, whereas high performance in the "against" class.
- Three main difficulties:

i) The inherent complexity of profiling the stance of ironic authors that employ stereotypes.
ii) the short size of the corpus.
iii) the imbalance between “in-favour” and “against” classes which made challenging the learning process.

Subtask take away

This task opens a new way to study ironic language to perpetuate stereotypes and constitutes a starting point for profiling authors who frame aggressiveness, toxicity and messages of hatred towards social categories such as immigrants, women and the LGTB+ community, using an implicit way to convey hate speech employing stereotypes.

Conclusions

- Several approaches to tackle the task:
  - Transformers (BERT-based), also combined with traditional representations and methods, obtained the highest results.

- Results:
  - Over 89% on average.
  - Best (99.44%): Yu et al. - BERT + CNN

- Error analysis:
  - False positives (non-irony spreaders as spreaders): 15.45%
  - False negatives (irony spreaders as non-spreaders): 9.18%

Conclusions

One of the main challenges of this task was to contemplate the use of stereotypes in a broad sense, that is, not focusing on a target group but considering those users who explain what happens in their environment by intensively using social categories.

Behind this theoretical approach there is the idea that prejudice is fundamentally a vision of the world that homogenizes people on the basis of their groups of origin or affiliation. A vision of the world that considers that these group affiliations are the main cause of the people's behaviours and could explain social or economic problems.

It is evident that to embrace stereotyping towards many social groups may have introduced a topic bias, although certainly when we analyse stereotypes towards a single group, the type of discourse changes if what is held is a stereotypical view of a group (certain social categories are brought up in order to present certain arguments). For example, gays are brought up in a moral discourse and immigrants are evoked in an economic or legal discussion, then it is important to take into account this association between target groups and topics.

- Corpus analysis illustrates that Ironic and non-Ironic users significantly differ not only in the use of Twitter elements but also in the indices used to characterise language, use of emotions, and communication styles, what could explain the high scores of the classifiers, and it opens the door to future research in order to characterise better the use of irony.

Conclusions

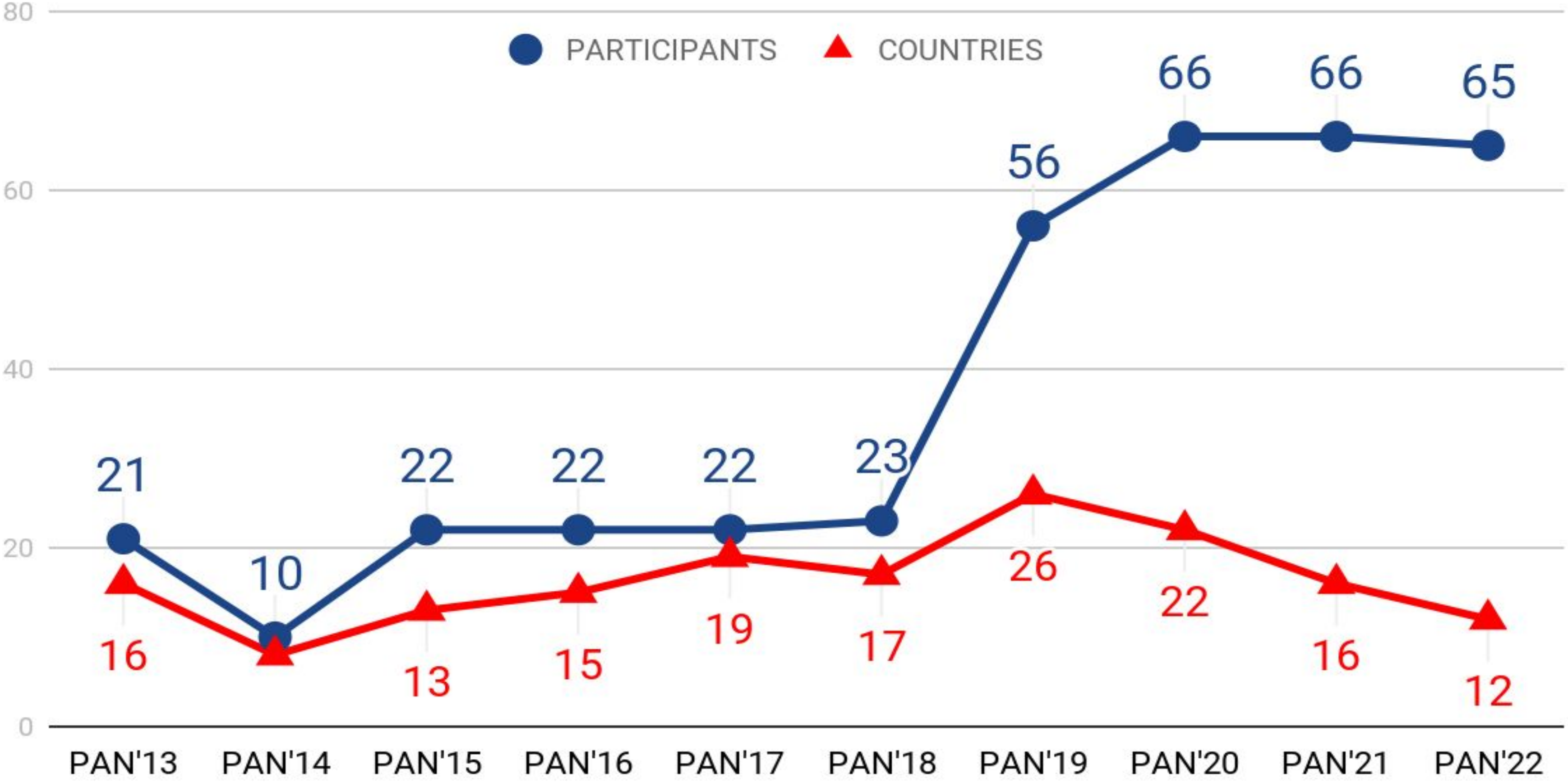
Looking at the results, the corpus analysis and the error analysis, we can conclude:

- It is feasible to automatically discriminate between Irony and non-Irony Spreaders with high precision
  - ...even when only textual features are used.

- Not only are the topics addressed by both types of users significantly different but also other elements such as the number of emojis they use, the number of users they mention, the number of hashtags they use, the number of URLs they share, their writing style, the emotions they convey or even their personality and communication style.

- We have to bear in mind false positives since 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
65
16
13
15
19
17
26
22
16
12
PAN'13
PAN'14
PAN'15
PAN'16
PAN'17
PAN'18
PAN'19
PAN'20
PAN'21
PAN'22

Industry at PAN (Author Profiling)

Organisation

🇪🇸 symanto psychology ai

Sponsors

🇩🇪 symanto psychology ai

This year, the winners of the task are:

- Wentao Yu, Benedikt Boenninghoff, and Dorothea Kolossa, Institute of Communication Acoustics, Ruhr University Bochum, Germany

PAN'23: Profile cryptocurrency influencers in social media from a low-resource perspective:

- Low-resource influencer profiling.
- Low-resource influencer interest identification.
- Low-resource influencer intent identification.

On behalf of the author profiling task organisers:

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