Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO).

Overview for PAN at CLEF 2022

Reynier Ortega-Bueno[1], Berta Chulvi[1,4], Francisco Rangel[2], Paolo Rosso[1] and Elisabetta Fersini[3]

[1]*Universitat Politècnica de València, Spain*
[2]*Symanto Research, Spain*
[3]*Università Degli Studi di Milano-Bicocca, Italy*
[4]*Universitat de València, Spain*

Abstract

This overview presents the Author Profiling shared task at PAN 2022. This year's task (IROSTEREO) focuses on determining whether the author of a Twitter feed is keen to spread irony and stereotypes. The main aim is to show the feasibility of automatically identifying potential Twitter users that spread stereotypes using indirect speech such as irony. For this purpose, a corpus with Twitter data in English has been provided. Altogether, the approaches of 64 participants have been evaluated. Moreover, a subtask on profiling stereotype stance at author level was also proposed in order to see if stereotypes have been employed by ironic authors to hurt the possible targets (e.g. immigrants, women, the LGTB+ community, etc.) or, on the contrary, to support them.

Keywords

Author profiling, Irony, Stereotypes, Social categories, Machine learning

1. Introduction

Language is, without doubt, one of the most creative skills among all mankind's cleverness. It is an extraordinary powerful machinery based on an idea of ingenious simplicity which allows us to compose out of twenty-five to almost forty sounds (e.g. Romance languages and English) an infinite variety of expressions. Beyond straightforward phonemes mix (grapheme in case of writing), which are constrained by lexical, syntactic and semantic rules, such expressions allow us to disclose to others its whole inside. Language does not only provide a straightforward way to communicate with each other by direct speech/writing, but also it enables indirect communication through creative and figurative language devices.

Irony is one of the most pervasive figurative device used in everyday communication and in social media platforms. Irony[1] implies the use of words that mean the opposite of what is really intended [1]. Usually, people use ironic speech/writing to express negative "private states"(sentiment, opinions, attitudes, beliefs, etc.) where the positive surface meaning differs from the implied one. This linguistic shift in meaning produced by ironic language endows

CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5-8, 2022, Bologna, Italy

CEUR Workshop Proceedings (CEUR-WS.org)

[1] The concept of irony is used in this work as an umbrella term for related phenomena such as sarcasm.

humans with a valuable resource to explore creativity in language and semantics. However, it simultaneously provides a tool that can indirectly and subtly mask language of hatred, offence and discrimination towards specific individuals or social groups. The work introduced by [2] addressed the issue of the hurtfulness of sarcasm in the content of social media. The authors found that ironic expressions of irony involve very negative emotions and sarcastic messages tend to be expressed with a more hurtful language, revealing the aggressive intention of the author towards the targeted victim.

In the context of interpersonal communication, it was introduced the problem of irony bias [3, 4, 5]. The authors investigated the role of verbal irony in the communication and maintenance of social stereotypes. They observed that irony is found more appropriated in situations in which stereotypes are violated than in situations in which social stereotypes are confirmed. A biased use of irony contributes to social stereotyping and may increase prejudice against minority groups. For that reason, it is crucial to detect, and if it is possible, contrast the diffusion of abusive, discriminatory, and stereotypic language also when it is disguised by figurative devices like irony and sarcasm.

Having previously focused on hate speech spreaders [6], at PAN'22 we have addressed the problem of profiling irony and stereotypes spreaders in social media, more specifically on Twitter. Special emphasis was given to those authors that employ irony to spread stereotypes, for instance, towards women, immigrants or the LGTB+ community. The goal is to classify authors as ironic or not, depending on their number of tweets with ironic content. This will allow for identifying possible stereotype spreaders on Twitter, as a first step towards preventing it. Our hypothesis is that users who do not spread irony and stereotypes may have a set of different characteristics compared to users who do. For example, they may use different linguistic patterns, writing style or affective information when they share posts compared to hate speech spreaders.

The remainder of this paper is organized as follows. Section 2 covers the state of the art on irony detection, author profiling, and stereotypes in language. Section 3 describes the corpus and the evaluation measures, and Section 4 presents the approaches submitted by the participants. Section 5 discusses the results achieved by the participants. Section 6 draws on analyses on the training dataset. Section 7 is devoted to the subtask of profiling stereotype stance at ironic author level. Finally, Section 8 presents the conclusions.

2. Related Work

The purpose of this section is to provide the theoretical background. We outline the relevant works in computational irony detection, author profiling and stereotypes in language.

2.1. Irony Detection

The problem of computationally irony detection has been investigated from different perspectives. Pioneer works had focused on the role of some surface linguistic features obtained from the text on its own such as n-grams, punctuation marks, part-of-speech tags, and simple syntactic patterns, among others [7, 8, 9, 10, 11]. Other works focused on some theoretical aspects of irony, such as unexpectedness, contradiction and opposition. Based on that, several features for capturing semantic ambiguity and polarity contrast have been studied [12, 13, 14, 15]. Similarly, some research agrees on the affective component behind ironic messages. Several approaches have stressed affective information for improving irony detection [16, 17, 18, 19]. Verbal irony is, without any doubt, a pragmatic phenomenon; hence, contextual and extra-linguistic information is crucial for its comprehension. In this direction, information concerning the context surrounding a given message has been used to determine whether a text has an ironic or sarcastic intention [20, 21, 22, 23].

Recently, deep learning based methods have attracted the focus of the research in several NLP tasks, including irony detection. In this direction, in [24] a pre-trained Robustly Optimized BERT Pre-training Approach (RoBERTa) [25] model was used to represent the sentences. After that, these were contextualized using a Recurrent Convolutional Neural Network (RCNN) to address irony and sarcasm detection. The authors in [26] proposed to use a Transformer architecture to contextualize pre-trained word embeddings. Specifically, they contextualized Word2Vec word embeddings, trained with several millions of tweets both for English and Spanish. This strategy, opposite to pre-trained Bidirectional Encoder Representations from Transformers (BERT), allows the system to be trained from in-domain representations using the same robust backbone architecture as BERT. From another point of view, the model introduced in [27], proposed strategies to improve irony detection by transferring knowledge from sentiment resources. For that, the authors proposed three different attentive Long Short Term Memory (attentive-LSTM) approaches that differ in the way of including the sentiment resources, either injecting the sentiment directly to the attention mechanisms or merging the output of different networks specialized on sentiment analysis and irony detection. In [28] an attentive-LSTM model was proposed for irony and satire detection in Spanish. The model takes advantage of three representations learned from, sentence-embedding, the BERT-based model and linguistic features. These representations were used to inform the proposed attentive-LSTM model to improve irony detection. In a similar fashion, in [2] a transformed-based system was introduced. The authors investigated the impact of hurtful and affective features on irony and sarcasm detection in Italian tweets.

From a multilingual point of view, most of the research carried out on irony detection has been done in English. Notwithstanding, there have been some efforts to investigate such figurative language device in other languages such as: Chinese [29], Czech [11], Dutch [10], French [30], Italian [31], Portuguese [7, 32], Spanish [33, 34], and Arabic [35, 36]. Even when the classification of a text as ironic or not has been widely studied from different perspectives, there are no references to computational works that attempt to profile authors who can be considered ironic and utilize this rhetorical figure to spread and perpetuate stereotypes on social media platforms.

2.2. Author Profiling

Pioneer researchers on author profiling focused on the analysis of blogs and formal texts [37, 38], based on Pennebaker's [39] theory. This theory connects the use of the language with the personality traits of the authors.

With the rise of social media, researchers proposed methodologies to profile the authors of posts where the language is more informal [40]. Since then, several approaches have been explored. For instance, [41] approached the age and gender identification problem with a second order representation which relates documents and user profiles. The authors of [42] proposed Low Dimensionality Statistical Embedding (LDSE), a statistical embedding which drastically reduces the dimensionality while seizing the whole vocabulary, and that has been commonly used as baseline in author profiling shared tasks.

Recently, the research community has focused on the usage of emotions and personality traits to address different problems, such as the Emograph graph-based approach enriched with topics and emotions [43]. Furthermore, due to the lack of large annotated corpora to train the new and powerful deep learning methods, the research community is focusing on using little or no training data to address the author profiling task [44].

During the past three editions, at PAN we focused on profiling users who spread harmful information, as well as profiling bots due to their key role in its propagation on Twitter. Concretely, in 2019 the goal was discriminating bots from humans [45], in 2020 identifying possible fake news spreaders [46], and in 2021 the focus was on profiling potential hate speech spreaders [6]. This year we aim at identifying potential spreaders of ironic contents, mainly when referring to special social categories or stereotypes.

2.3. Stereotypes in Language

Stereotypes are generally defined as a set of widespread beliefs that are associated with a group category [47]. This is the theoretical assumption of the Stereotype Content Model (SCM) [48] very popular in computational linguistic approaches to stereotype detection. SCM states that two dimensions persist in social cognition when people are making sense of individuals or groups: perceived warmth (trustworthiness, friendliness) and competence (capability, assertiveness). Supporting this approach, most of the attempts to study stereotypes in Computational Linguistics have focused on a particular target group such as gender [49, 50], ethnic minorities [51], religion [52], immigrants [53, 54, 55] and age [56]. Most of them use a word embeddings representation and rely on the association of attributes to a social group. The common goal has been to identify which stereotypical beliefs are associated to each particular group, introducing bias in large language models, which are increasingly used in AI applications.

However, what most research overlooks is the fact that before of this description of group in terms of concrete traits, a previous homogenization of the group must be done. This homogenization of diversity is at the base of the over-generalization that allows the success of the stereotypical reasoning. As Lippman argued in his seminal work [57] about stereotypes, this cognitive process that disregards the variability of the real world occurs because "we do not first see and then define, we define first and then see". The original idea of stereotype, as Lippmann defines it, relies more on a cognitive process that disregards diversity into a group or into a particular event than in the content itself of a particular stereotype.

Following this idea, we start from the premise that a vision of the word in terms of social categories is previous to the use of a stereotype, that is to say, is previous to this cognitive process that assumes that a singular individual has some characteristics simply based on their perceived membership in the group. In this sense, in this task of author profiling, we have tried to operationalize the idea that some people share a vision of the world that intensively uses social categories to describe and explain reality, a prejudiced mentality that systematically privileges a worldview in terms of homogenous social groups and undervalues the internal diversity of any social group.

Few works have approached the study of stereotypes affecting more than a target group. The authors in [58] create StereoSet, a large-scale natural English dataset to measure stereotypical biases in four domains: gender, profession, race, and religion. They contrast both stereotypical bias and language modelling ability of popular models like BERT, GPT2, RoBERTa, and XLNET, showing that these models exhibit strong stereotypical biases. Recently, Sap and colleagues [59] approach also the presence of several target groups in the Social Bias Frame, a new conceptual formalism that aims to model the pragmatic frames in which people project social bias and stereotypes onto others. To support this research, they developed the Social Bias Inference Corpus (SBIC) with 150,000 structured annotations of social media posts covering 34,000 implications about social groups. For example, in front of a sentence as "If cameras do really add ten pounds, do Africans really exist?", annotators from Amazon Mechanical Turk indicate whether or not: (i) the post is offensive, (ii) the intent is to offend, and (iii) it contains sexual content. Only if annotators indicate potential offensiveness they answer the group implication question: who is referred to/targeted by this post? Two possible answers were: (i) yes, this could be offensive to a group and (ii) no, this is just an insult to an individual or a non-identity-related group of people. If the post targets or references a demographic group, annotators select or write which group is referenced. For each selected group, they then write two to four stereotypes that are used in this post; for the given example, annotators write as stereotype: "Africans are all starving". Finally, workers are asked whether they think the speaker is part of one of the minority groups referenced by the post. From 16,739 instances in SBIC, 8,167 refer to a group of people in the field of "target minority". On the basis of this work we have constructed the IROSTEREO corpus as it is explained in Section 3.1.

3. Evaluation Framework

The purpose of this section is to introduce the technical background. We outline the construction of the corpus, introduce the performance measures and baselines, and describe the software submissions.

3.1. IROSTEREO Corpus

In this section, we describe the methodology followed to build the corpus, introduce the taxonomy and the stereotype categories, explain the retrieval of the tweets and the annotation process, and finally give some statistics of the obtained corpus.

3.1.1. Taxonomy and Stereotype Categories

To build the IROSTEREO corpus we examine the "target minority" field of the SBIC by [59] which has 150,000 structured annotations of social media posts covering 34,000 implications about social groups. We identify 600 unique labels that could be considered a social category in SBIC. We define a social category following a long tradition of research in social psychology [60] [61] which considers that a social group exist when two or more persons define themselves as members of the group and when their existence is recognised by at least one other person. Sap et al. [59] classify the groups referenced in seven categories: (1) body (2) culture (3) disabled (4) gender (5) race (6) social and (7) victims.

In order to focus specifically on stereotypes as the expression of a prejudice against certain social categories that are often the object of an ironic and hurtful discourse, we create a more granular taxonomy to classify the 600 labels in 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 and (17) white people. As keywords to retrieve the tweets, we use the labels associated to groups only from categories 5 to 14 of the taxonomy.

3.1.2. Tweet Retrieval and Annotation Process

The Twitter API was used to retrieve tweets with two conditions: (i) tweets that contain the hashtag irony or sarcasm and at least one of the labels included in categories 5 to 14 of the taxonomy and (ii) the same labels about social groups but without irony or sarcasm. Users with more cases in classes (i) and (ii) were identified and the tweets that accomplish these two conditions were downloaded. The annotators had to identify ironic tweets and tweets that use stereotypes among this set of users. To identify irony, the annotators were asked to mark the tweets where the user "express the opposite of what was saying as a disguised mockery". If a user had more than five ironic tweets, it was labelled as ironic.

To identify the use of stereotypes, annotators were asked to check if the social categories present in the tweets were used to refer to a social group by associating them with a homogenising image of the category. For example, they talk about gays or Muslims in general, as if they were all similar, and could be well described with that word. If a user had more than five tweets containing a stereotyped image of a group, the user was labelled as a user who utilises stereotypes. Positive examples of classes 1 (users that express irony without stereotypes), 2 (non-ironic users that use stereotypes) and 3 (users that express irony and use stereotypes) were selected and 200 tweets from their timeline were downloaded. To find the non-ironic and non-stereotype class (4) the lexicon used in the three previous classes was analysed in order to reduce topic bias. Moreover, tweets should not contain the labels of social categories associated to stereotypes.

The annotation process was carried in two steps. During the first one, data were annotated by two independent annotators. The inter-annotator agreement (IAA) between the first two annotators was 0.7093. During the second one, those instances where a disagreement exists, we asked for a third annotation to solve it. Moreover, the second annotation was done also to check that class 4 does not contain irony samples.

3.1.3. Corpus Statistics

Table 1 presents the statistics of the corpus that consists of 600 authors for English language, completely balanced between the two classes (ironic and non-ironic), and with a 66/33 balance between users employing stereotypes or not for each class. For each author, we retrieved via the Twitter API their timeline and sampled 200 tweets. We have split the corpus into training and test sets, following a proportion of 70/30 for training and testing respectively.

Table 1
Number of authors in the PAN-AP-22 IROSTEREO corpus distributed between the two classes, Ironic vs Non-Ironic, and within each class, distributed between users who use stereotypes vs. users who do not use stereotypes.

	Ironic			Non-Ironic			
Set	**Stereotypes**	**Non-Stereo.**	**Total**	**Stereotypes**	**Non-Stereo.**	**Total**	**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

3.2. Performance Measure

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.

3.3. Baselines

As baselines to compare the performance of the participants with, we have selected:

- *RF + char 2-grams* character $bigrams$ and Random Forest classifier.
- *LR + word 1-grams* Bag of Words (BOW) with Logistic Regression classifier.
- *LSTM+Bert-encoding* We represent each tweet in the profile utilising pretrained Bert-base model. Later, we fed an LSTM with these vectors as input.
- *LDSE* [42]. 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 vs no-irony spreader. The distribution of weights for a given document should be closer to the weights of its corresponding category.

3.4. Software Submissions

Similar to previous year[2], we asked for software submissions. Within software submissions, participants submitted executables of their author profiling software instead of just the output of their software on a given test set. For the software submissions, the TIRA experimentation platform was employed [62, 63], which renders the handling of software submissions at scale as simple as handling run submissions. Using TIRA, participants deploy their software on virtual machines at our site, which allows us to keep them in a running state [64].

[2] This year we also have allowed some users to directly sent us their prediction files as well as their software for us to reproduce their systems.

4. Overview of the Participating Systems

This year, 65 teams participated in the author profiling shared task and 33 of them submitted the notebook paper. We analyse their approaches from three perspectives: preprocessing, features used to represent the authors' texts and classification approaches.

4.1. Preprocessing

In order to prevent bias towards some URLs, user mentions or hashtags, the corpus was provided with these elements already masked. In the same vein, some participants cleaned other Twitter-specific elements such as RT, VIA, and FAV[3] reserved words [65, 66, 67, 68, 69, 70, 71, 72, 73], as well as emojis and other non-alphanumeric characters [74, 71, 70, 75, 73], numbers [73, 69, 74] or punctuation signs [69, 67, 76, 73, 65, 70]. Several participants lower-cased the texts [65, 76, 77, 78, 79, 80, 69, 74, 81], removed stop words [74, 73, 65], or stemmed or lemmatised the terms [82]. Some users also removed infrequent terms or meaningless ones [65, 79, 70]. The authors in [83] carried out a X^2 test which combined with Pointwise Mutual Information (PMI) and TF-IDF was used to select the most contributing words to the representation, or the authors in [75] who added the labels (I/NI) to the end of each tweet. Finally, the authors in [82] used GloVe as a pre-trained embedding matrix to filter out features.

4.2. Features

The participants have used a high variety of different features and their combinations, albeit we can group them into the following main groups: *(i)* n-grams; *(ii)* stylistics; *(iii)* personality and emotions; and *iv)* deep learning-based such as embeddings and transformers.

Regarding n-grams, a high variety of them have been used by several authors, in most of the cases in combination with other representations. Mainly, character n-grams [81, 74], word n-grams [83, 72, 81, 74] (sometimes weighted with TF-IDF [84]), and syntactic n-grams (e.g. POS) [74].

The authors of [74] have combined different types of n-grams with sentiment and emotions, as well as hateful content (aggressive, hateful and targeted), while the authors of [84] have combined stylistic features such as average vocabulary size, average number of tokens, average tweet length, average number of hashtags, mentions and URLs, average number of emojis, or the ratio between lowercased and uppercased words, the LiX score, TF-IDF unigrams, TF-IDF profanity, TF-IDF emojis, Part of Speech (POS) tags count, sentiment analysis and punctuation signs.

Different transformers have been also widely used to extract features. For instance, BERT [85, 73, 79, 75, 86, 66, 67, 69, 71], SBERT [87], BERTweet [70] or combining them with other feature extraction methods. For example, BERT and Twitter RoBERTa with LM HateXPlain [88] fine-tuned with the HatEval dataset [89], SBERT with emojis [87], psychometrics, emotions and irony with SBERT [90], BERT with TF-IDF n-grams [91, 92], or SBERT with graph-based and one-hot embeddings [65].

[3] RT is the acronym for *retweet*; VIA is a way to give the authorship to a user (e.g., "via @kicorangel"); and FAV stands for *favourite*.

The authors of [93] have combined different stylometric features such as lexicon-based, social media jargon or POS with static embeddings (FastText) and contextual embeddings obtained with BERT and RoBERTa. Similarly, the authors of [68] have combined different n-grams with stylistic features based on lexicons with sentence transformers.

Convolutional Neural Networks (CNNs) have been also used to extract features [78]. The authors of [82] combined a CNN with TF-IDF unigrams and a Bidirectional Gated Recurrent Unit (BiGRU) to represent the authors' texts. The authors of [80] have used a TextVectorizer to extract their features while word embeddings have been used by the authors of [76, 94]. The authors of [95] have combined sequence probabilities and n-grams with GPT2 and DistilGPT2.

The authors of [86] combined a semantic representation obtained with a transformer with punctuation signs and auxiliary words representations. Similarly, the authors of [90] have obtained ironic-, contextual- and psychometric-related features with transformers that had been fine-tuned with datasets annotated with sentiment and emotions from the Kaggle competition[4].

Finally, the authors of [96] approached the task by identifying irony at individual tweet level. To that end, they have combined three types of features: *i)* structural features such as punctuation marks, length of words, part-of-speech labels, Twitter marks, semantic similarity, etc.; *ii)* sentiment words by applying different lexical resources such as AFINN, Hu&Liu, and SentiWordNet; and *iii)* fine-grained emotions, by means of emotional lexicons such as EmoLex, EmoSenticNet, ANEW, Dictionary of Affect in Language, and SenticNet, among others.

4.3. Approaches

Most participants have used traditional approaches, mainly Random Forest (RF) [72, 96, 83, 84], Logistic Regression (LR) [91, 74], Bayes (NB) [95], Multilayer Perceptron (MLP) [73], Gradient Booster Classifier (GBC) [90], or k-Nearest Neighbours (k-NN) [86].

Ensembles of classifiers have been also used by various authors. For example, Support Vector Machine (SVM) and RF with a hard voting classifier [81], SVM with a Grading Boosting Classifier [90], or SVM, RF and LR with soft- and hard-voting ensemble [79]. Some participants have combined traditional approaches with deep learning ones through stacking ensembles with Logistic Regression with a meta-learner and SVM, Naive Bayes and Decision Trees (DT), together with CNN [77], and a meta-learner with SVM, DT, Naive Bayes and CNN [80].

Deep learning has been widely used to approach this year task, mainly CNNs [85, 76, 78], Graph Convolutional Neural Networks (GCNN) [65], Linear Feed Forward Networks [87], as well as combinations such as Bidirectional Long Short Term Memory (BiLSTM) and CNN [94] or just fully-connected networks [82, 93, 89]. Some participants used AutoML (AutoKeras) [70] and AutoGluon [71, 69] to automate the selection of the classifier.

With respect to transformer-based approaches, BERT and some of its variants have been the most used ones, usually combined with other approaches. For instance, BERT with Decision Rules [75], BERT with SVM, MLP, Gaussian Naive Bayes and RF [86], BERT with CNN, LSTM and attention layer [92], BERT and DistilBERT with RF and SVM [68], or just several BERT with a voting classifier [66, 67].

[4] https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text

5. Evaluation and Discussion of the Results

In this section, we present the results of the shared task, as well the analysis of the most common errors made by the teams.

5.1. Overall Ranking

In Table 2, the overall performance (in terms of accuracy) of the participants is presented. The top-ranked participants approached the task as follows. The overall best result (99.44%) has been obtained by Yu *et al.* [85] with a BERT feature-based CNN model. The second best result has been achieved by Tahaei *et al.* [87] with a combination of SBERT and emojis. The *two ex aequo* third best performing teams, respectively, used a Multilayer Perceptron trained with features extracted from a pre-trained BERT model[5], and a Random Forest fed with unigrams pre-selected with several techniques such as Chi^2, PMI, and TF-IDF with the aim to maximise the probability difference of each feature for each class [83].

Table 2
Overall accuracy of the submission to the task on profiling irony and stereotypes spreaders on Twitter.

	TEAM	ACCURACY
1	wentaoyu [85]	0.9944
2	harshv [87]	0.9778
3	edapal	0.9722
3	ikae [83]	0.9722
5	JoseAGD [93]	0.9667
5	Enrub	0.9667
7	fsolgui	0.9611
7	claugomez [92]	0.9611
9	AngelAso	0.9556
9	alvaro [86]	0.9556
9	xhuang [95]	0.9556
9	toshevska	0.9556
9	tfnribeiro_g [84]	0.9556
14	josejaviercalvo	0.9500
14	taunk [72]	0.9500
14	your	0.9500
14	PereMarco	0.9500
14	Garcia_Sanches	0.9500
19	pigeon	0.9444
19	xmpeiro	0.9444
19	marcosiino [77]	0.9444
19	dingtli	0.9444
19	moncho	0.9444
19	yifanxu	0.9444
19	yzhang [71]	0.9444
19	longma	0.9444
	LDSE	0.9389
27	missino [80]	0.9389
27	badjack	0.9389
27	sgomw	0.9389
27	wangbin	0.9389 [70]
27	caohaojie [66]	0.9389
32	lwblinwenbin [67]	0.9333
32	xuyifan [69]	0.9333
32	dirazuherfa [96]	0.9333
32	Los Pablos	0.9333
32	Metalumnos	0.9333
37	narcis	0.9278
37	stm [78]	0.9278
37	huangxt233 [95]	0.9278
40	lzy [79]	0.9222
40	avazbar	0.9222
40	fragilro	0.9222
40	whoami	0.9222
40	Garcia_Grau	0.9222
45	hjang [68]	0.9167
45	nigarsas	0.9167
45	fernanda [81]	0.9167
45	Hyewon	0.9167
49	zyang [94]	0.9056
50	giglou [65]	0.9000
50	sulee	0.9000
52	ehsan.tavan [90]	0.8889
53	rlad	0.8778
54	balouchzahi [74]	0.8722
	RF + char bigrams	0.8610
55	manexagirrezabalgmail	0.8500
	LR + word unigrams	0.8490
56	tamayo [89]	0.8111
57	yuandong [76]	0.7500
	LSTM+Bert-encoding	0.6940
58	G-Lab	0.6778
58	AmitDasRup [91]	0.6778
60	Alpine_EP	0.6722
61	Kminos	0.6667
62	castro [86]	0.6389
63	castroa	0.5833
64	sokhandan	0.5333
64	leila [73]	0.5333

Table 3
Statistics on the accuracy.

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

[5] The participants did not submit their working notes but sent us a brief description of their system.

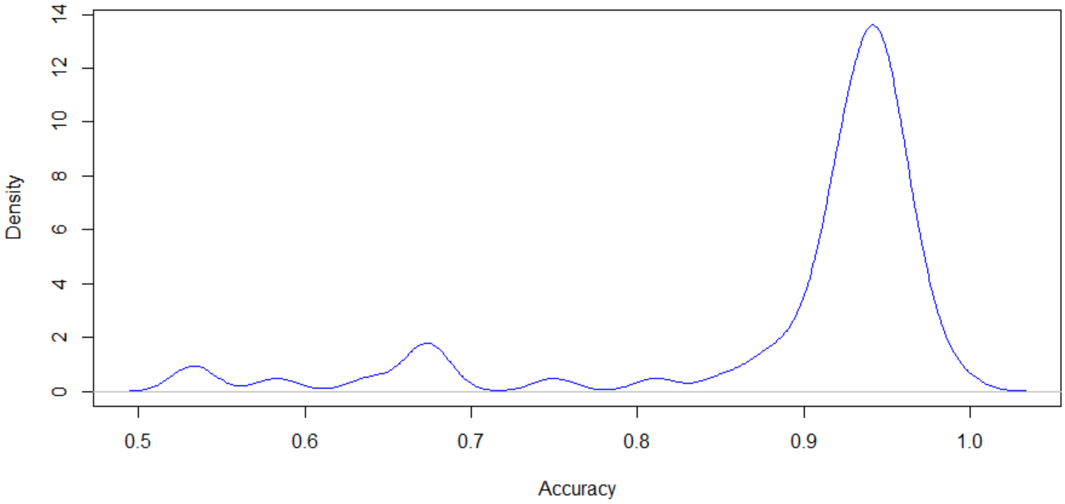

Figure 1: Density of the results in terms of accuracy.

As can be observed in Figure 1 and Table 3, the results do not follow a normal distribution (p-value = 2.2e-16) when we consider all of them. There are several outliers on the bottom side of the distribution, as can be also seen in Figure 2. When getting rid of the outliers (from the left), the top performing systems do follow the normal distribution (p-value | acc>0.85 = 0.1382), allowing the usage of the t-student test for the comparison of the significance of their differences. In these regards, the best performing team is not significantly better than the second and third ones ($z_c = 1.3451$ and $z_c = 1.6435$). Indeed, statistically significances appear with respect to the fifth best performing team ($z_c = 2.1418$).

5.2. Error Analysis

We have aggregated all the participants' predictions for irony vs non-irony spreaders, except baselines, and plotted the confusion matrix in Figure 3. It can be seen that the error is higher in the case of false positives (from non-irony to irony spreaders): 15.45% vs. 9.18%. This higher number of false positives is something to be investigated further in future research since it may introduce a bias towards the ironic class.

6. Corpus Analysis

With the aim to study how the authors of the different classes (irony vs non-irony) use the language, in this section, we analyse in detail: (i) the most commonly used topics per class; (ii) the usage of Twitter elements such as the number of words, hashtags, mentions and shared URLs; (iii) their writing style; (iv) the emotions they convey; (v) and their psychographics and

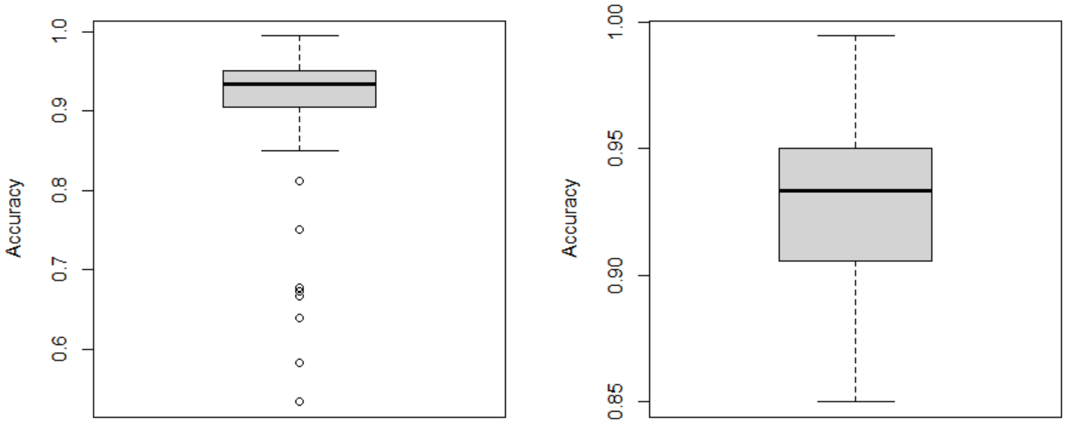

Figure 2: Distribution of results in terms of accuracy. The figure on the left represents all the systems. The figure on the right removes the outliers.

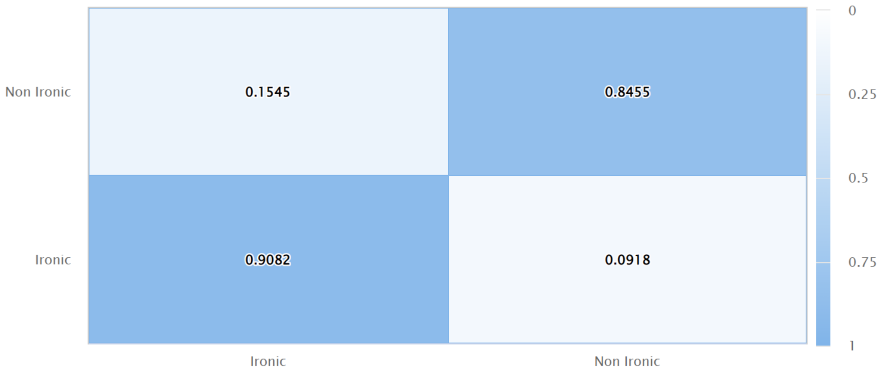

Figure 3: Aggregated confusion matrix for irony vs non-irony spreaders in English.

communication styles.

6.1. Topic-based Analysis

As described in Section 3.1, we collected the users by querying Twitter APIs with a list of keywords associated with irony and social categories, and the annotators labelled the proper usage of ironic language and social categories. However, the keyword-based data collection process might have introduced a bias in the corpus regarding the topics they cover.

The problem of topic bias has been analysed on hate speech corpora. Particularly, in the works [97, 98] the authors computed statistical scores in order to determine the correlation between the words and hate speech microblogs. In this work, we perform a two-fold analysis in our corpus similar to the one of [98]:

i Determining the set of unique words in each class and analysing how this vocabulary impacts the learning process;

ii Determining the set of words that are highly polarized according to the indexes introduced in [98].

We used the Polarized Wiredness Index (PWI) which takes into account how polarized the words are in each class in the corpus (irony and non-irony). PWI compares the relative frequency of a word as it occurs in the subset of a labelled dataset identified by one value of the label against its complement. Let us consider an annotated corpus $C = \{(d_1, l_1), (d_2, l_0), ..., (d_n, l_1)\}$ where $d_i = (w_1, w_2, w_3, ..., w_m)$ represents the i^th document in C, and w_j the words in d_i, with $i = 1, ..., |C|$ and $l_i \in [0, 1]$. The PWI of w_j w.r.t. the label l_0 is the ratio of the relative frequency of w_j in the subset $d_i \in C : l_i = l_0$ over the relative frequency of w_j in the complement subset $d_i \in C : l_i = l_1$.

$$PWI(w_j, l) = \frac{N_l(w_j)/T_l)}{N_{\hat{l}}(w_j)/T_{\hat{l}})}$$

Where $N_l()$ and $N_{\hat{l}}()$ represent the frequency of the term w_j in the class $l = 0$ and in its complement $l = 1$, respectively. T_l and $T_{\hat{l}}$ represent the total count of words in the class l (irony) and in its complement (non-irony), respectively.

Regarding the first point, we identified 1,379 words which only appear in one class[6]. In the irony class, we found 334 unique terms, whereas, in the class non-irony, we identified 1,045 unique terms. With the aim of investigating how this vocabulary may impact the learning process, we trained an SVM and RF classifiers considering as features the words in this vocabulary. As a result, the RF model achieved an Acc=0.8763 and the SVM an Acc=0.8817.

Later on, we analyse if even when the words are in both classes, their representativeness in each one is biased. For that, the PWI index was computed for the words in the corpus. Table 4 illustrates the highest-ranking words (25 words) according to their PWI in both classes.

Table 4

List of words from the IROSTEREO corpus with highest Polarized Weirdness Index (PWI) for No-Irony class (left column), and highest PWI for the Irony class (right column)

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*

From the first column, it can be noticed that in the Non-Irony class there are high-PWI words related to ethnic (e.g. *aboriginals, ados, africans, americas, bipoc, charlottesville*, etc.). Looking

[6] It is essential to note that their frequency is low in the whole corpus

at the high-PWI words of the second column, the most characteristic words in the Irony class are related to politics (e.g., *conway, djt, du30, duterte*) and religion (e.g., *antizionist, bhakts, brampton, dowry and cathie*), but also to some stylistic elements like emojis. In order to analyse how these polarized words may have impacted the learning process, we select the 100 most polarized words in each class. We train a couple of SVM and an RF classifiers, considering as features the words in this reduced vocabulary. As a result, the RF model achieved an Acc= 0.8833 and the SVM an Acc=0.8333. As can be noticed, the models achieve high classification scores. This analysis confirms that these words may have introduced a topic bias. However, we have also found out that stylistic aspects like emojis are also more used in the Irony class, which means that the topic bias might have been introduced by the strategy adopted to collect the data, or may have been also caused by a latent bias induced by the authors themselves.

6.2. Twitter Elements Analysis

In this subsection, we analyse the usage of Twitter elements such as hashtags, user mentions and URLs, as well as the average number of words used by the different types of users. The usage of these elements by the authors of the corpus is not normally distributed according to the Kolmogorov-Smirnov test (p<.001 for all the cases, both in the training and test sets). Therefore, we have performed the Mann-Whitney test to compare the corresponding distributions of these variables in the two classes (ironic vs. non-ironic users) in the two sets (training and test). The Mann-Whitney test relies on scores being ranked from lower to higher. Therefore, the group with the lowest mean rank is the group with the greatest number of lower scores and vice versa. For all the variables, the Mann-Whitney test is significant in both sets (see Table 5). In Table 5 we observe that non-ironic users (Mdn= 18.14) employ more number of words than ironic users (Mdn= 12.79). Both classes also differ in the use of typical Twitter elements. As we can see in Figure 4, ironic people use more hashtags, more mentions, and fewer URLs than non-ironic ones.

Table 5
Statistics about differences between ironic and non-ironic users in the usage of Twitter elements.

ELEMENT	CLASS	N	MEDIAN	MEAN RANK	MANN-WHITNEY U	p-value
N. WORDS	IRONIC	300	12.79	268.62	35,437	<.001
	NON-IRONIC	300	18.14	332.38		
HASHTAGS	IRONIC	300	0.24	353.68	60,954	<.001
	NON-IRONIC	300	0.14	247.32		
MENTIONS	IRONIC	300	0.88	335.81	55,593	<.001
	NON-IRONIC	300	0.54	265.19		
URLS	IRONIC	300	0.19	269.63	35,740	<.001
	NON-IRONIC	300	0.22	331.37		

6.3. Language Style Analysis

To analyse whether there are differences between ironic and non-ironic users in their language style, we apply the POS-tagging FreeLing[7] to each tweet of every user. A total of 33 morphological features are used to represent texts. A score of each morphological feature is computed as

[7] https://nlp.lsi.upc.edu/freeling/index.php/

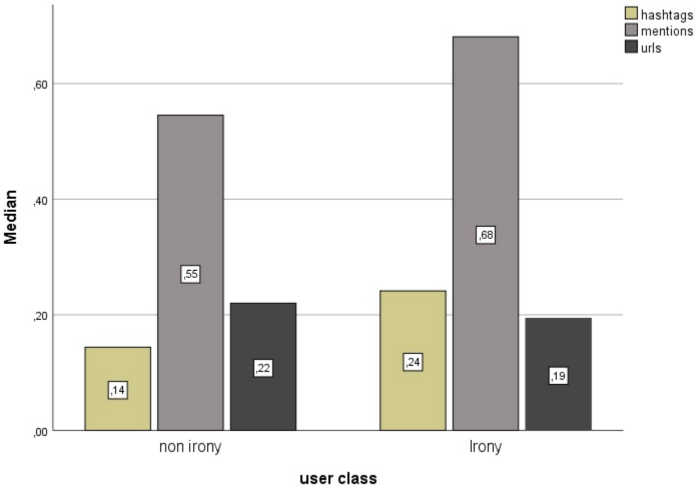

Figure 4: Differences between ironic and non-ironic users in their usage of Twitter elements

the percentage of this feature over the total words of the tweet. Then, the average score for each user in every morphological feature is calculated and normalized. With these scores, the categorical vs narrative index used by authors in previous research is calculated. Inspired in the work of Nisbett [99], the categorical versus narrative index is computed as a simple algorithm: nouns + adjectives + prepositions - verbs - adverbs - personal pronouns. Positive values in this index express more categorical style of language and negative values more narrative style. Categorical style is used to express ideas and concepts, whereas narrative is used to tell stories.

The scores of this index are not normally distributed according to the Kolmogorov-Smirnov test (p<.001 in training and test data), then we perform the Mann-Whitney Test to compare the distributions of this variable in the two classes: ironic users vs non-ironic users. The Mann-Whitney test is significant in both sets of data (test and training), then we offer the statistics for the entire corpus in Figure 5. As we can see, non-ironic users (Mdn=0.71) utilize significative more than ironic users (Mdn= -0.99; U=15,834; p<.001) a categorical language style. Ironic users utilise more a narrative style. It is interesting to notice that these features are topic-agnostic, and we can conclude that, regardless of the topic, there are significant differences in the way ironic and non-ironic users employ language.

6.4. Emotions Analysis

The new Dictionary of Affect in Language [100] is used to test if the ironic and non-ironic users differ in the expression of emotions. The new Dictionary of Affect in Language (DAL for short) is an instrument designed to measure the emotional meaning of words and texts. It compares individual words to a word list of 8,742 words that were originally rated by 200 naïve volunteers along three dimensions: activation (active vs. passive), imaginary (easy vs. difficult to imagine), and pleasantness (unpleasant vs. pleasant).

The scores of the three dimensions are not normally distributed according to the Kolmogorov-

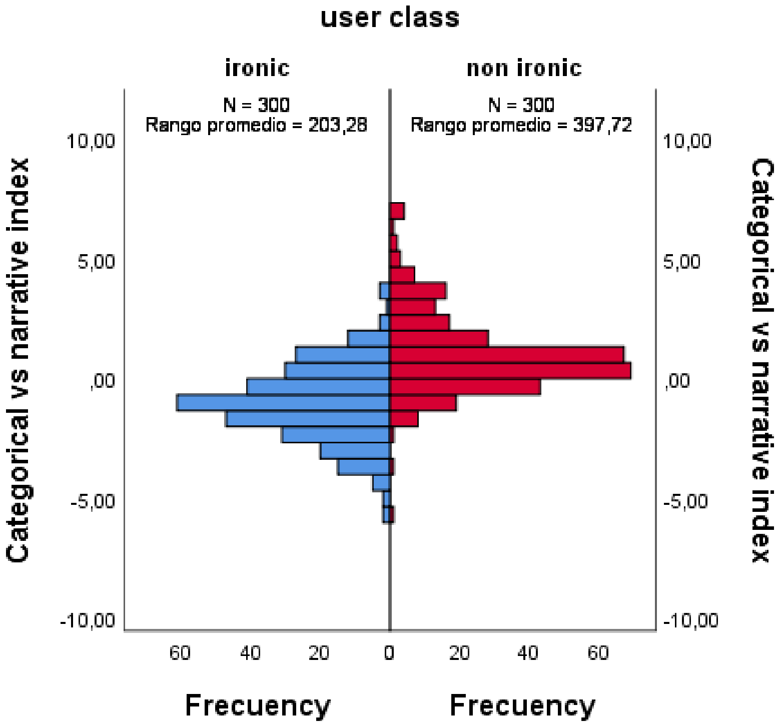

Figure 5: Differences between ironic and non-ironic users in the categorical vs narrative index

Smirnov test (p<.001 for all of them in training and test data), then we perform the Mann-Whitney Test to compare the scores distributions of these three dimensions in the two classes: ironic users vs non-ironic users. The Mann-Whitney test is significant for the three dimensions in both sets of data, then we offer the statistics for the entire corpus (Table 6). As we can see in Figure 6, non-ironic users present higher scores in the three emotional dimensions than ironic ones.

6.5. Communication Styles

We use the Symanto API[8] to obtain the users' personality type and communication styles [101]. The personality type refers to the way the person behaves in a specific interaction from the emotional vs rational point of view. Regarding communication styles, it is composed of four traits: (i) *action-seeking*, defined as direct or indirect requests, suggestions, and recommendations that

[8] https://rapidapi.com/collection/symanto-symanto-default-apis

Table 6
Statistics about the differences between ironic and non-ironic users in the use of emotions

DAL DIMENSION	CLASS	N	MEDIAN	MEAN RANK	MANN-WHITNEY U	p-value
ACTIVATION	Ironic	300	0.31	229.65	23,744	p<.001
	Non-ironic	300	0.33	371.35		
IMAGINERY	Ironic	300	0.28	197.89	14,216	p<.001
	Non-ironic	300	0.31	403.11		
PLEASANTNESS	Ironic	300	0.33	238,83	26,499	p<.001
	Non-ironic	300	0.36	362,17		

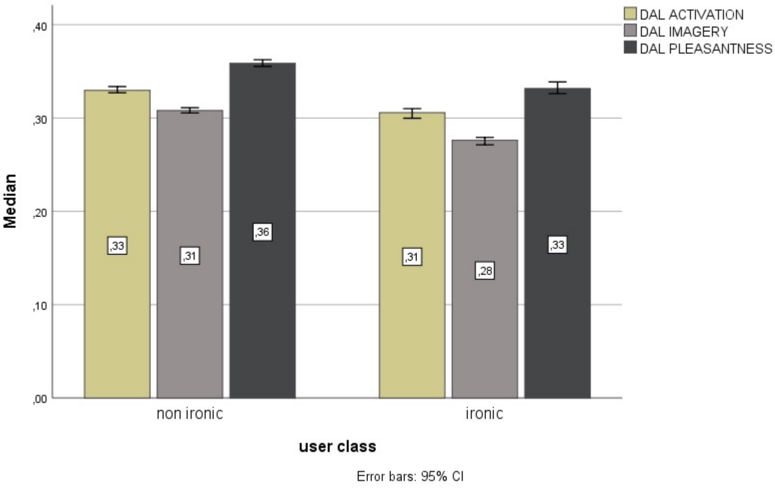

Figure 6: Differences between ironic and non-ironic users in the use of emotions (DAL dimensions)

expect action from other people; (ii) *fact-oriented*, where the user utilises factual and objective statements; (iii) *self-revealing*, when the users share personal information or experiences; and (iv) *information-seeking*, defined as direct or indirect questions searching for information. For each of these traits, the Symanto API returns a value between 0 and 1, representing the confidence for the person to belong to some part of the continuous between the two extremes of the trait, for example, somewhere between completely emotional and completely rational.

The data of all these variables are not normally distributed according to the Kolmogorov-Smirnov test (p<.01 for all of them in training and test data, except for information-seeking where the p-value in training is p<.001 and in the test is .09). We perform the Mann-Whitney Test to compare the distributions of these variables in the two classes: ironic users vs non-ironic users in the training and in the test sets. For all variables, except action-seeking styles, the Mann-Whitney test is significant in the training and in the test data. Table 7 illustrates the statistics for the entire corpus. As we can see in Figure 7, ironic users use less the fact-oriented style and more a self-revealing and information-seeking style. In the personality type measure, the non-ironic users are more emotional and less rational than the ironic users.

Table 7
Statistics about the differences between ironic and non-ironic users in communication styles and personality type

INDEX	CLASS	N	MEDIAN	MEAN RANK	MANN-WHITNEY U	EXACT SIG.
Action-seeking	Ironic	300	0.082	299.70	44,761	.910
	Non-ironic	300	0.089	301.30		
Fact-oriented	Ironic	300	0.172	237.16	25,999	p<.001
	Non-ironic	300	0.238	363.84		
Self-revealing	Ironic	300	0.702	332.32	54,545	p<.001
	Non-ironic	300	0.675	268.68		
Information-seeking	Ironic	300	0.092	373.33	66,849	p<.001
	Non-ironic	300	0.063	227.67		
Rational vs. emotional	Ironic	300	0.488	271.30	36,240	p<.001
	Non-ironic	300	0.491	329.70		

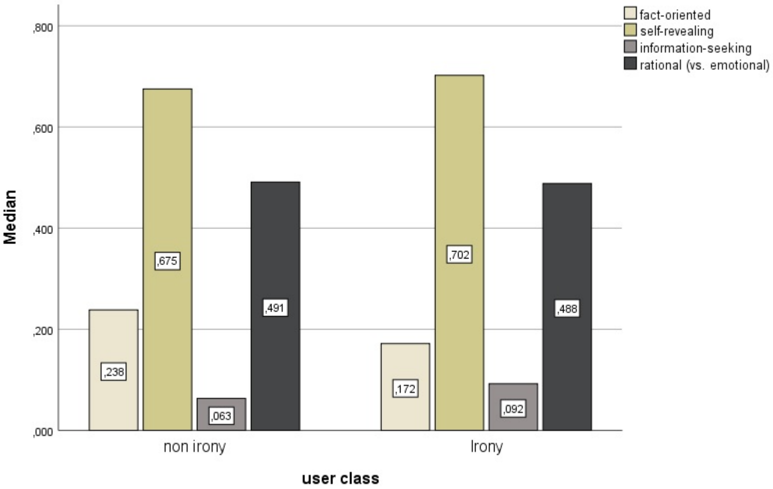

Figure 7: Differences between ironic and non-ironic users in communication styles and personality types

7. Profiling Stereotype Stance of Ironic Authors

In this section, we aim to investigate the usage of irony to refer to stereotypes via the analysis of the authors' stance toward the targets. In fact, stereotypes may have been employed by ironic authors to hurt the targets (e.g. immigrants, women, the LGBT+ community, etc.) or to somehow support them. This subtask aims at detecting the stance of how stereotypes are used by ironic authors, whether in favour or against the target. Therefore, given the subset of ironic authors that employed stereotypes in some of their tweets, the goal is to detect their overall stance. In the four examples below, it can be observed how ironic messages can be used to support or hurt the target.

- *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]**

7.1. IROSTEREO-Stance Corpus

For creating the IROSTEREO-Stance corpus, we selected those authors that were annotated as ironic and spreaders of stereotypes in IROSTEREO. Later, we performed a third annotation process on this data: for each author, only the tweets marked as ironic and using stereotypes in the IROSTEREO corpus were annotated with their stance. We did not provide any kind of guidelines for the annotation. Instead, we asked the annotators to rely on their own perspectives on whether the tweets are in favour or against the mentioned social category. The overall stance of an author is considered "*in-favour*" if the majority of the annotated tweets in her profile support the targets; in the other case, it is considered as "*against*". The overall stance of each ironic author that used stereotypes was initially annotated by two independent annotators. The IAA between the first two annotators was 0.645. Then, for those ironic authors where a disagreement existed, we asked third annotator for another annotation. Finally, a dataset with 58 ironic authors "*in-favour*" and 142 "*against*" was obtained. The distribution in training and test is showed in Table 8.

Table 8
Number of authors in the IROSTEREO-Stance corpus distributed between the two classes, In-Favour vs Against

SET	IN-FAVOUR	AGAINST	TOTAL
Training	46	94	140
Test	12	48	60

7.2. Experimental Results

The performance of the systems is evaluated using the macro averaged F1 measure (F_Macro), although we also analyse the F1-measure per class to study more in depth +the behaviour of the systems (F1_A and F1_F for the "against" and "in-favour" class, respectively). We have evaluated three baselines in the profiling stereotype stance subtask:

- *RF + char 3-grams* character $trigrams$ and Random Forest.
- *SVM + word 2-grams* $bigrams$ of words with Support Vector Machine.
- *LDSE method* [42]

In this subtask, we did not constrain the number of runs that a team could summit and the 7 teams submitted 15 runs in total. All results achieved by each team and the baselines are shown in Table 9.

Table 9
F1_Macro of the participating systems in the subtask of profiling stereotype stance of ironic authors.

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

Most of the teams tested on the IROSTEREO-Stance corpus the systems previously submitted to the IROSTEREO task. The models submitted by *dirazuherfa's team* [96] employed emotIDM, an emotion-based approach, that comprises three groups of features for representing the tweets: (i) structural features: punctuation marks, length of words and chars, part-of-speech labels, Twitter marks (i.e., hashtags, mentions, etc.), and semantic similarity; (ii) sentiment features: an overall value of polarity is calculated in terms of how many positive or negative words a tweet contains (the sentiment intensity of each word was considered); (iii) emotions features: information regarding emotions from several lexicons. Moreover, an oversampling method was applied to address the imbalance in the training set. The runs differ from each other in the subset of features and in the classification model. Run1 and run4 considered all features in emoIDM combined with a 7-NN and 5-NN, respectively. In run2 and run3, only the structural features were considered for training a 5-NN and a 3-NN classifier, respectively. The *toshevska's team* trained a deep graph convolutional neural network (HinSAGE) to classify user nodes. For that, a heterogeneous graph was created. It comprises three types of nodes: user nodes, tweet nodes, and nodes, and three types of edges: user-tweet, tweet-word, and word-word. The system used by the *JoseAGD's team* [93] relied on four-faced representations. The feature representations are based on linguistic features from UMUTextStats, non-contextual sentence embeddings from FastText, contextual embeddings from BERT and contextual embeddings from RoBERTa. Later, a fully connected neural network was trained. Run1 submitted by the *tamayo's team* [89] used a prototype creation strategy for representing the profiles. Firstly, the tweets in the profile are encoded employing a pretrained RoBERTa-based model. Based on these representations, the profile is split into two groups, one where the tweets are strongly related, according to their inner similarity, and another group with more heterogeneous information. The representation of the profile is the sum of the tweet's encoding from the former group. For classifying a new author, a KNN method was applied. In *run2* the tweets are encoded using three transformers models: BERT-base, Twitter-RoBERTa-base and LM HateXplain. The profile was modelled using a Spatial Graph Convolutional Neural Network, and a fully connected dense network was used as the classifier. The *run1* and *run2* submitted by *AmitDasRup's team* [91] used BERT combined with the TF-IDF representation, and the prediction was made by a logistic regression classifier. The runs differ in the parameters of TF-IDF used to build the vocabulary. The *taunk's team* [72] represented tweets using Bag of Word and TF-IDF weighting. Later, the profiles were built as the sum of the tweet vectors. Based on this representation of the profiles, several shallow machine learning models were trained. The authors experimented with Random Forest, Support Vector Machines, K Nearest Neighbors, Logistic Regression, and XGBoost. The best result was achieved using the SVM model. Finally, *fernanda's team* [81] proposed an ensemble method based on a hard voting scheme. Three distinct representations: char n-grams, word n-grams and Out of Vocabulary (OOV) were built. After that, SVM and RF were used as base classifiers, and their prediction was aggregated in the voting schema.

As it can be observed in Table 9 the results achieved by all participants are moderated. Five runs reached an F1_Macro>0.50, and no participants outperformed the LDSE baseline. Also, it can be noticed that the systems had a low performance in the “in-favour" class, whereas high F1 scores are achieved in the class “against”. We hypothesize that the three main problems that have been faced by the participating systems are: i) the inherence complexity of profiling the stance of ironic authors that employ stereotypes, ii) the short size of the IROSTEREO-Stance corpus; and iii) the imbalance between “in-favour” and “against” classes which made challenging the learning process. Although the results were quite modest, this task opened 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.

8. Conclusions

In this paper, we have presented the results of the 10th International Author Profiling Shared Task at PAN 2022, hosted at CLEF 2022. The participants had to discriminate on Twitter between irony and no-irony spreaders. The provided data cover the English language.

The participants used different features to address the task, mainly: *i)* n-grams; ii) stylistics; *iii)* personality and emotions; and *iv)* deep learning-based representations such as embeddings and transformers. Concerning machine learning algorithms, the most used ones were combinations and ensembles of different traditional algorithms such as SVM, Logistic Regression and Random Forest with deep learning techniques such as Fully-Connected Neural Networks, CNN, LSTM and Bi-LSTM, and transformer-based ones, mainly BERT and its variations.

The best result (99.44%) has been obtained with a BERT feature-based CNN model. The second best result (97.78%) has been achieved with a combination of SBERT and emojis, and the two *ex aequo* third best results (97.22%), respectively, with a Multilayer Perceptron trained with features extracted from a pre-trained BERT model, and a Random Forest fed with unigrams pre-selected with several techniques of feature selection.

The error analysis shows that the highest confusion is towards irony spreaders (false positives) with almost double number of errors (15.45% vs 9.18%), which requires further research to prevent systems to bias their predictions.

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.

Another conclusion derived from the corpus analysis is that ironic and non-ironic users differ significantly not only in the use of Twitter elements but also in the indices used to characterise language, use of emotions, and communication styles, which could explain the high scores obtained by the classifiers. These consistent differences in style open the door to future research in order to characterize better the use of irony.

Looking at the results, the corpus analysis and the error analysis, we can conclude that: *i)* it is feasible to automatically discriminate between irony and non-irony spreaders with high accuracy; *ii)* 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; *iii)* we have to bear in mind false positives since they are almost double than false negatives, and misclassifications might lead to ethical or legal implications [102].

Acknowledgments

The work of the third author has been partially funded by CDTI under grant IDI-20210776, IVACE under grant IMINOD/2021/72, and grant PLEC2021-007681 funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR. The work of the fourth author was in the framework of the DeepPattern research project funded by the Generalitat Valenciana (PROMETEO/2019/121).

References

[1] A. Reyes, P. Rosso, On the Difficulty of Automatically Detecting Irony: Beyond a Simple Case of Negation, in: Knowledge and Information Systems, vol. 40, issue 3, pp. 595-614, 2014.

[2] S. Frenda, A. T. Cignarella, V. Basile, C. Bosco, V. Patti, P. Rosso, The Unbearable Hurtfulness of Sarcasm, Expert Systems with Applications 193 (2022) 116398. URL: https://

www.sciencedirect.com/science/article/pii/S0957417421016870. doi:https://doi.org/10.1016/j.eswa.2021.116398.

[3] C. Burgers, C. J. Beukeboom, Stereotype Transmission and Maintenance Through Interpersonal Communication: The Irony Bias, Communication Research 43 (2016) 414–441. doi:10.1177/0093650214534975.

[4] C. J. Beukeboom, C. Burgers, How Stereotypes are Shared through Language: A Review and Introduction of the Social Categories and Stereotypes Communication (SCSC) Framework, Review of Communication Research 7 (2019) 1–37. doi:10.12840/issn.2255-4165.017.

[5] C. J. Beukeboom, C. Burgers, Seeing Bias in Irony: How Recipients Infer Speakers' Stereotypes from their Ironic Remarks about Social-Category Members, Group Processes and Intergroup Relations 23 (2020) 1085–1102. doi:10.1177/1368430219887439.

[6] F. Rangel, G. L. De la Peña Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling Hate Speech Spreaders on Twitter Task at PAN 2021, in: CLEF (Working Notes), 2021, pp. 1772–1789.

[7] P. Carvalho, L. Sarmento, M. J. Silva, E. de Oliveira, Clues for Detecting Irony in User-generated Contents: Oh...!! it's "so easy" ;-), in: Proceedings of the 1st International Conference on Information Knowledge Management Workshop on Topic-Sentiment Analysis for Mass Opinion, 2009, pp. 53–56.

[8] D. Davidov, O. Tsur, A. Rappoport, Semi-supervised Recognition of Sarcastic Sentences in Twitter and Amazon, in: Proceedings of the Fourteenth Conference on Computational Natural Language Learning, CoNLL '10, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 107–116.

[9] R. González-Ibáñez, S. Muresan, N. Wacholder, Identifying Sarcasm in Twitter: A Closer Look, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT '11, Association for Computational Linguistics, Portland, Oregon, 2011, pp. 581–586.

[10] F. Kunneman, C. Liebrecht, M. van Mulken, A. van den Bosch, Signaling Sarcasm: From Hyperbole to Hashtag, Information Processing & Management 51 (2015) 500 – 509.

[11] T. Ptáček, I. Habernal, J. Hong, Sarcasm Detection on Czech and English Twitter, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 213–223.

[12] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, R. Huang, Sarcasm as Contrast between a Positive Sentiment and Negative Situation, in: Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), 2013, pp. 704–714.

[13] F. Barbieri, H. Saggion, Automatic Detection of Irony and Humour in Twitter, Proceedings of the Fifth International Conference on Computational Creativity (2014) 155–162.

[14] F. Barbieri, H. Saggion, Modelling irony in Twitter: Feature Analysis and Evaluation, Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014 (2014) 4258–4264.

[15] C. V. Hee, Can Machines Sense Irony ?, Ph.D. thesis, Universiteit Gent, 2017.

[16] D. I. Hernández Farías, V. Patti, P. Rosso, Irony Detection in Twitter: The Role of Affective Content, ACM Trans. Internet Technol. 16 (2016) 19:1–19:24. doi:10.1145/2930663.

[17] D. I. Hernández Farías, J.-M. Benedí, P. Rosso, Applying Basic Features from Sentiment

Analysis for Automatic Irony Detection, in: R. Paredes, J. S. Cardoso, X. M. Pardo (Eds.), Pattern Recognition and Image Analysis, volume 9117 of *Lecture Notes in Computer Science*, Springer International Publishing, Santiago de Compostela, Spain, 2015, pp. 337–344. doi:10.1007/978-3-319-19390-8_38.

[18] F. Barbieri, H. Saggion, F. Ronzano, Modelling Sarcasm in Twitter, a Novel Approach, in: Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 50–58.

[19] A. Reyes, P. Rosso, T. Veale, A Multidimensional Approach for Detecting Irony in Twitter, Language Resources and Evaluation 47 (2013) 239–268.

[20] D. Bamman, N. A. Smith, Contextualized Sarcasm Detection on Twitter, in: Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, AAAI, Oxford, UK, 2015, pp. 574–577.

[21] A. Khattri, A. Joshi, P. Bhattacharyya, M. Carman, Your Sentiment Precedes You: Using an Author's Historical Tweets to Predict Sarcasm, in: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Lisboa, Portugal, 2015, pp. 25–30.

[22] B. C. Wallace, D. K. Choe, E. Charniak, Sparse, Contextually Informed Models for Irony Detection: Exploiting User Communities, Entities and Sentiment, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 1035–1044.

[23] A. Ghosh, T. Veale, Fracking Sarcasm using Neural Network, in: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, San Diego, California, 2016, pp. 161–169. URL: http://www.aclweb.org/anthology/W16-0425.

[24] R. A. Potamias, G. Siolas, A. G. Stafylopatis, A Transformer-based Approach to Irony and Sarcasm Detection, Neural Computing and Applications 32 (2020) 17309–17320. URL: https://doi.org/10.1007/s00521-020-05102-3. doi:10.1007/s00521-020-05102-3.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[26] J. Á. González, L. F. Hurtado, F. Pla, Transformer based Contextualization of Pre-trained Word Embeddings for Irony Detection in Twitter, Information Processing and Management 57 (2020) 1–15. URL: https://doi.org/10.1016/j.ipm.2020.102262. doi:10.1016/j.ipm.2020.102262.

[27] S. Zhang, X. Zhang, J. Chan, P. Rosso, Irony Detection via Sentiment-based Transfer Learning, Information Processing & Management 56 (2019) 1633 – 1644. doi:10.1016/j.ipm.2019.04.006.

[28] R. Ortega-Bueno, P. Rosso, J. E. M. Pagola, Multi-view Informed Attention-based Model for Irony and Satire Detection in Spanish Variants, Knowledge-Based Systems 235 (2022) 107597.

[29] Y.-j. Tang, H.-H. Chen, Chinese Irony Corpus Construction and Ironic Structure Analysis, in: Proceedings of COLING 2014, the 25th International Conference on Computational

Linguistics, Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 1269–1278.

[30] J. Karoui, F. Benamara, V. Moriceau, N. Aussenac-Gilles, L. Hadrich-Belguith, Towards a Contextual Pragmatic Model to Detect Irony in Tweets, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, 2015, pp. 644–650.

[31] C. Bosco, V. Patti, A. Bolioli, Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT, IEEE Intelligent Systems 28 (2013) 55–63.

[32] U. B. Corrêa, L. Coelho, L. Santos, L. A. de Freitas, Overview of the IDPT Task on Irony Detection in Portuguese at IberLEF 2021, Procesamiento del Lenguaje Natural 67 (2021) 269–276.

[33] G. Jasso López, I. Meza Ruiz, Character and Word Baselines Systems for Irony Detection in Spanish Short Texts, Procesamiento del Lenguaje Natural 56 (2016) 41–48. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5285.

[34] R. Ortega-Bueno, F. Rangel, D. Hernández Farıas, P. Rosso, M. Montes-y Gómez, J. E. Medina Pagola, Overview of the Task on Irony Detection in Spanish Variants, in: Proceedings of the Iberian languages evaluation forum (IberLEF 2019), co-located with 34th conference of the Spanish Society for natural language processing (SEPLN 2019). CEUR-WS. org, volume 2421, 2019, pp. 229–256.

[35] J. Karouia, F. B. Zitoune, Veronique Moriceau, SOUKHRIA: Towards an Irony Detection System for Arabic in Social Media, in: 3rd International Conference on Arabic Computational Linguistics, ACLing 2017, Association for Computacional Linguistic (ACL), Dubai, United Arab Emirates, 2017, pp. 161–168.

[36] B. Ghanem, J. Karoui, F. Benamara, V. Moriceau, P. Rosso, IDAT at FIRE2019: Overview of the Track on Irony Detection in Arabic Tweets, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, 2019, pp. 10–13.

[37] S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, Gender, genre, and writing style in formal written texts, Text & Talk 23 (2003) 321–346.

[38] M. Koppel, S. Argamon, A. R. Shimoni, Automatically categorizing written texts by author gender, Literary and linguistic computing 17 (2002) 401–412.

[39] J. W. Pennebaker, M. R. Mehl, K. G. Niederhoffer, Psychological Aspects of Natural Language Use: Our Words, our Selves, Annual review of psychology 54 (2003) 547–577.

[40] J. D. Burger, J. Henderson, G. Kim, G. Zarrella, Discriminating Gender on Twitter, Technical Report, MITRE CORP BEDFORD MA BEDFORD United States, 2011.

[41] A. P. López-Monroy, M. Montes-y Gómez, H. J. Escalante, L. Villasenor-Pineda, E. Villatoro-Tello, INAOE's Participation at PAN'13: Author Profiling Task, in: CLEF 2013 evaluation labs and workshop, 2013.

[42] F. Rangel, P. Rosso, M. Franco-Salvador, A Low Dimensionality Representation for Language Variety Identification, in: In 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing'16. Springer-Verlag, LNCS(9624), 2018, pp. 156–169.

[43] F. Rangel, P. Rosso, On the Impact of Emotions on Author Profiling, Information processing & management 52 (2016) 73–92.

[44] M. Chinea-Rios, T. Müller, G.-L. De-la-Peña Sarracén, F. Rangel, M. Franco-Salvador, Zero and Few-Shot Learning for Author Profiling, in: Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2022, p. 333–344. URL: https://doi.org/10.1007/978-3-031-08473-7_31. doi:10.1007/978-3-031-08473-7_31.

[45] F. Rangel, P. Rosso, Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling, in: CLEF 2019 Labs and Workshops, Notebook Papers, 2019.

[46] F. Rangel, A. Giachanou, B. Ghanem, P. Rosso, Overview of the 8th Author Profiling Task at PAN 2019: Profiling Fake News Spreaders on Twitter, in: CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, 2020.

[47] J. Dovidio, M. Hewstone, P. Glick, V. Esses, Prejudice, Stereotyping, and Discrimination: Theoretical and Empirical Overview, The SAGE Handbook of Prejudice, Stereotyping and Discrimination (2010) 3–28. doi:10.4135/9781446200919.n1.

[48] S. T. Fiske, Stereotype Content: Warmth and Competence Endure, Current Directions in Psychological Science 27 (2018) 67–73. URL: https://doi.org/10.1177/0963721417738825. doi:10.1177/0963721417738825. arXiv:https://doi.org/10.1177/0963721417738825, pMID: 29755213.

[49] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, A. Kalai, Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (2016).

[50] E. Fersini, P. Rosso, M. E. Anzovino, Overview of the Task on Automatic Misogyny Identification at IberEval 2018, in: IberEval@SEPLN, 2018.

[51] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes, Proceedings of the National Academy of Sciences 115 (2018) E3635–E3644. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1720347115. doi:10.1073/pnas.1720347115. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1720347115.

[52] A. Abid, M. Farooqi, J. Zou, Large Language Models Associate Muslims with Violence, Nature Machine Intelligence 3 (2021) 461–463. doi:10.1038/s42256-021-00359-2.

[53] M. Sanguinetti, G. Comandini, E. Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task, 2020.

[54] J. Sánchez-Junquera, B. Chulvi, P. Rosso, S. P. Ponzetto, How Do You Speak about Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants, Applied Sciences 11 (2021). URL: https://www.mdpi.com/2076-3417/11/8/3610. doi:10.3390/app11083610.

[55] J. Sánchez-Junquera, P. Rosso, M. M. y Gómez, B. Chulvi, Masking and BERT-based Models for Stereotype Identification, Procesamiento del Lenguaje Natural 67 (2021) 83–94. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6379.

[56] K. C. Fraser, S. Kiritchenko, I. Nejadgholi, Extracting Age-Related Stereotypes from Social Media Texts, in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3183–3194. URL: https://aclanthology.org/2022.lrec-1.341.

[57] W. Lipmann, Public Opinion, New York:Harcourt Brace, 1922.

[58] M. Nadeem, A. Bethke, S. Reddy, StereoSet: Measuring Stereotypical Bias in Pretrained Language Models, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 5356–5371. URL: https://aclanthology.org/2021.acl-long.416. doi:10.18653/v1/2021.acl-long.416.

[59] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social Bias Frames: Reasoning about Social and Power Implications of Language, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5477–5490. URL: https://aclanthology.org/2020.acl-main.486. doi:10.18653/v1/2020.acl-main.486.

[60] H. Tajfel, A. A. Sheikh, R. C. Gardner, Content of Stereotypes and the Inference of Similarity between Members of Stereotyped Groups, Acta Psychologica, 22 (1964) 191–201.

[61] R. Brown, Prejudice. Its Social Psychology, Wiley-Blackwell, 2010.

[62] T. Gollub, B. Stein, S. Burrows, Ousting ivory tower research: Towards a web framework for providing experiments as a service, in: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 1125–1126.

[63] T. Gollub, B. Stein, S. Burrows, D. Hoppe, TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments, in: 2012 23rd International Workshop on Database and Expert Systems Applications, IEEE, 2012, pp. 151–155.

[64] T. Gollub, M. Potthast, A. Beyer, M. Busse, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, Recent Trends in Digital Text Forensics and its Evaluation, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2013, pp. 282–302.

[65] H. B. Giglou, M. Rahgouy, A. Rahmati, T. Rahgooy, C. D. Seals, Profiling Irony and Stereotype Spreaders with Encoding Dependency Information using Graph Convolutional Network, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[66] H. Cao, Z. Han, Z. L. Zhenwei Mo, Z. Xiao, Z. Li, L. Kong, A Multi-Model Voting Ensemble Classifier based on BERT for Profiling Irony and Stereotype Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[67] W. Lin, Z. Han, J. Zhang, Z. Li, G. Cao, J. Yu, L. Kong, A BERT-based Model for Profiling Irony and Stereotype Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[68] H. Jang, Lexicon-Based Profiling of Irony and Stereotype Spreaders-Notebook for PAN at CLEF 2022, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[69] Y. Xu, H. Ning, Profiling Irony and Stereotype Spreaders on Twitter with BERT, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[70] B. Wang, H. Ning, Notebook for PAN at CLEF 2022:Profiling Irony and Stereotype Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[71] Y. Zhang, H. Ning, Irony and Stereotype Spreaders Detection using BERT-large and AutoGulon, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[72] D. T. adn Sagar Joshi, V. Varma, Profiling Irony and Stereotype Spreaders on Twitter based on Term Frequency in Tweets, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[73] L. Hazrati, A. Sokhandan, L. Farzinvash, Profiling Irony Speech Spreaders on Social Networks Using Deep Cleaning and BERT, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[74] S. Butt, G. S. Fazlourrahman Balouchzahi, A. Gelbukh, CIC@PAN: Simplifying Irony Profiling using Twitter Data, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[75] L. K. Jian Qin, J. Huang, Use Pre-trained Models and Multi-classifier Voting Methods to Identify the Ironic Authors on Twitter, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[76] L. M.-Z. T. Dong Yuan, Wenyin Yang, Q. Lao, Analysis of Irony and Stereotype Spreaders Based On Convolutional Neural Networks, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[77] I. T. Marco Siino, M. L. Cascia, T100: A Modern Classic Ensembler to Profile Irony and Stereotype Spreaders, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[78] M. S. Stefano Mangione, G. Garbo, Improving Irony and Stereotype Spreaders Detection using Data Augmentation and Convolutional Neural Network, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[79] L. Zengyao, H. Zhongyuan, An Ensemble Machine Learning Classifier for Profiling Irony and Stereotype Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[80] D. Croce, D. Garlisi, M. Siino, An SVM Ensembler Approach to Detect Irony and Stereotype Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[81] M. F. A. Herold, D. C. Castro, User Profiling: Voting Scheme, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[82] D. L. Haolong Ma, Y. Sun, Profiling Irony and Stereotype Spreaders on Twitter Using

TF-IDF and Neural Network, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[83] C. Ikae, UniNE at PAN-CLEF 2022: Profiling Irony and Stereotype Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[84] Y. Nikolova, K. Hano, T. Ribeiro, Irony Stereotype Spreader Detection using Random Forests, Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO) 2022, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[85] B. B. Wentao Yu, D. Kolossa, BERT-based Ironic Authors Profiling, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[86] A. Rodriguez, M. Barroso, Profiling Irony and Stereotype Spreaders on Twitter: PAN Shared Task (IROSTEREO) 2022, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[87] P. B.-N. S. Narjes Tahaei, Harsh Verma, S. Bergler, Identifying Authors Using Irony or Spreading Stereotypes with SBERT and Emojis, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[88] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A Benchmark Dataset for Explainable Hate Speech Detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 14867–14875.

[89] R. Labadie, D. Castro, Graph-Based Profile Condensation for Users Profiling, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[90] E. Tavan, M. Najafi, R. Moradi, Identifying Ironic Content Spreaders on Twitter using Psychometrics, Contextual and Ironic Features with Gradient Boosting Classifier, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[91] A. Das, N. Raychawdhary, G. Dozier, C. D. Seals, Irony Spreading Author Profiling on Twitter using Machine Learning: A BERT-TFIDF based Approach, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[92] C. Gómez, D. Parres, BERT Sentence Embeddings in different Machine Learning and Deep Learning Models for Author Profiling applied to Irony and Stereotype Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[93] J. A. G. Díaz, M. Ángel Rodríguez-García, F. García-Sánchez, R. Valencia-Garcia, UMUTeam at IROSTEREO: Profiling Irony and Stereotype spreaders on Twitter combining Linguistic Features with Transformers, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[94] W. Y.-Q. L. Zexian Yang, Li Ma, Z. Tan, A Intelligent Detection Method for Irony and Stereotype Based on Hybird Neural Networks, in: CLEF 2022 Working Notes, CEUR-WS.org, 2022.

[95] X. Huang, Profiling Irony and Stereotype Spreaders with Language Models and Bayes' Theorem, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[96] D. I. Hernández Farías, M. Montes-Y-Gómez, Exploiting Affective-based Information for Profiling Ironic Users on Twitter, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[97] M. Wiegand, J. Ruppenhofer, T. Kleinbauer, Detection of Abusive Language: The Problem of Biased Datasets, in: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 602–608.

[98] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review, Language Resources and Evaluation 55 (2021) 477–523.

[99] R. E. Nisbett, K. Peng, I. Choi, A. Norenzayan, Culture and Systems of Thought: Holistic Versus Analytic Cognition, Psychological review 108 (2001) 291.

[100] C. M. WHISSELL, Chapter 5 - THE DICTIONARY OF AFFECT IN LANGUAGE, in: R. Plutchik, H. Kellerman (Eds.), The Measurement of Emotions, Academic Press, 1989, pp. 113–131. URL: https://www.sciencedirect.com/science/article/pii/B9780125587044500116. doi:https://doi.org/10.1016/B978-0-12-558704-4.50011-6.

[101] S. Štajner, S. Yenikent, M. Franco-Salvador, Five Psycholinguistic Characteristics for Better Interaction with Users, in: 2021 8th International Conference on Behavioral and Social Computing (BESC), IEEE, 2021, pp. 1–7.

[102] F. Rangel, P. Rosso, On the Implications of the General Data Protection Regulation on the Organisation of Evaluation Tasks, Language and Law= Linguagem e Direito 5 (2019) 95–117.