Does anyone see the irony here? Analysis of perspective-aware model predictions in irony detection

Simona Frenda[1,2,*], Soda Marem Lo[1], Silvia Casola[1], Bianca Scarlini[3], Cristina Marco[3], Valerio Basile[1] and Davide Bernardi[3]

[1]Computer Science Department, University of Turin, Turin
[2]aequa-tech srl, Turin
[3]Alexa AI, Amazon Development Centre Italy, Turin

Abstract

In the framework of *perspectivism*, analyzing how people perceive pragmatic phenomena, like irony, is relevant for deeply understanding the different points of view, and for creating more robust perspective-aware models. This paper presents a linguistic analysis of irony perception in 11 perspectivist models. Each model is trained on annotations by crowd-sourcing workers different in gender, age, and nationalities. Due to the sparsity of the dataset, we examine the texts classified as ironic and not-ironic by these perspectivist models, and identify linguistic patterns that all perspectives associate with irony. To our knowledge, we are the first to also provide evidence for the different linguistic patterns perceived as ironic by a specific perspective. For example, models trained on data annotated by American and Australian annotators are more inclined to classify a text as ironic when it includes a negative sentiment, while models trained on data annotated by the youngest annotators are particularly influenced by words related to immoral behaviors.
***Warning**: This paper could contain content that is offensive or upsetting for the reader.*

Keywords

Irony Detection, Irony Interpretation, Perspectivism, Linguistic Analysis

1. Introduction

The use of supervised learning is the core of several areas of Artificial Intelligence, including Natural Language Processing (NLP). Models that leverage this learning paradigm are strictly dependent on either automatically-produced datasets, i.e., silver data, or manually-curated ones, i.e., gold standards. In the contest of human-made annotations, the standard approach determines the final annotation by resolving the disagreement of multiple annotators, e.g., through majority voting. Recent research trends offer an alternative take and show that flattening the disagreement of several annotators can discard valuable information [1, 2].

Some of these trends go by the name of *perspectivist approaches*. According to these lines of research, the discrepancies of different annotators can be exploited to model different points of view (*perspectives*) on a specific task [3]. This is especially important when the task is highly subjective, such as that of identifying irony [4]. While some linguistic patterns are linked to this phenomenon by a majority of people [5], irony tends to be closely related to the cultural and personal background of those who interpret it [6, 7].

In this paper, we investigate the perception of irony in different segments of the English-speaking population. We focus, in particular, on two research questions (RQ):

- RQ_1: what are the common linguistic triggers for irony interpretation, regardless of perspectives?
- RQ_2: what are the linguistic patterns typical of each perspective?

To answer these questions, we exploited EPIC (English Perspectivist Irony Corpus) [8], a disaggregated English corpus for irony detection, containing 3,000 pairs of *Posts-Replies* from Twitter and Reddit, along with the demographic information of each annotator.

Inspired by [9], and in continuity with [8], we grouped annotators in 11 different *perspectives*: self-identified female and male, age-based groups (boomers, generation X, generation Y and generation Z), and country-based groups. Then, reproducing the experiments of [8], we created 11 perspective-aware models and obtained their predictions on the same set of instances.

We do so to perform a quantitative and qualitative analysis of the common and specific linguistic patterns (affective, offensive, syntactic, and lexical) that activate the ironic interpretation of a text for each population segment. We leveraged the models' knowledge to predict

2nd Workshop on Perspectivist Approaches to NLP at ECAI 2023
*Corresponding author.
simona.frenda@unito.it (S. Frenda); sodamarem.lo@unito.it (S. M. Lo); silvia.casola@unito.it (S. Casola); scarlini@amazon.it (B. Scarlini); marcocri@amazon.it (C. Marco); valerio.basile@unito.it (V. Basile); dvdbe@amazon.it (D. Bernardi)
© 2023 Copyright © 2023 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

the labels on the TEST SET, and performed a linguistic analysis on this portion of the corpus to compare the predicted perception of each social group on the same content. In fact, since instances are annotated on average by 5 annotators, they do not necessarily contain labels for all demographic traits and perspectives. For example, an instance can be annotated by workers from Generation GenY and GenZ only, and lack labels from annotators of the older generations.

By comparing the relevance of different linguistic features for the perspectivist models, we are able, firstly, to confirm the importance – for all perspectives – of some specific features known to be of high impact in previous works [5]; secondly, we show that some patterns are perspective-specific.

For instance, we found that the models trained on female, generation Y, Australian, and American perspectives tend to recognize irony especially when the texts express negative sentiment. The Irish perspective seems to be amused by the emotional contrast in the texts. The male perspectivist model, instead, seems to be more sensitive to the recognition of irony when texts contain insults explicitly related to crimes or immoral behaviors, professions, and animals. A similar difference is also visible in the dimension of age, where words related to female genitalia appear relevant in the decision for Generation X; in contrast, the youngest generations (i.e., Y and Z) are more influenced by words related to crimes and immoral behaviors. Models trained on the perspectives of boomers and Indians are sensitive to specific syntactic patterns.

These analyses shed light on the different perceptions of irony by different population segments. While we found common patterns that are independent of languages and perspectives, attention to different points of view is needed especially for creating user-centered applications and for making them explainable.

This paper is organized as follows. In Section 2, we present an overview of previous works related to the analysis of linguistic features and strategies for expressing irony, focusing on a multilingual and multiperspective approach to the phenomenon. In Section 3 we describe the EPIC corpus, used to perform the source-independent (Section 4.1) and source-dependent (Section 4.2) analyses on the patterns that drive the interpretation of our perspective-aware models. Finally, Section 5 is dedicated to the discussion and conclusive observations on our results.

2. Related Work

Literature about irony detection has explored the contribution of several linguistic features within classical and neural architectures (using golden standard datasets): syntactic [10], stylistic [11], pragmatic [12], semantic [13], and affective [14, 15, 16] ones. Despite the clear impact of some of these features on irony detection, the general cognitive mechanisms that activate irony regardless of language and domain are still being studied [17, 5, 18, 19].

The authors of [5] conducted an exhaustive linguistic analysis on three Twitter datasets annotated for the irony detection task in French, Italian, and English. They looked for specific ***linguistic strategies*** used for expressing irony: analogy, metaphor, hyperbole/exaggeration, euphemism, rhetorical question, oxymoron, paradox, and other elements such as false assertion, context shift, situational irony, or ***specific markers*** (emoticons, negations, patterns of discourse, hashtags labelling the presence of humour, intensifiers, punctuation, false propositions, elements of surprise, modality, quotations, opposition, capital letters, personal pronouns, interjections, comparison, named entities, report verbs, expression of opinion, urls). Oxymorons, false assertion, and situational irony have been confirmed as triggers for irony in Italian tweets also by the authors of [20], who analysed the predictions obtained in the context of the IronITA shared task [21]. Unlike other languages, ellipsis and apostrophes stand out for Spanish [22].

Another common trait for irony detection from the multilingual perspective is the role played by affective information. For example, the authors of [14] showed how pleasantness, imagery, activation, and negative sentiment have a discriminative power in classifying ironic and non-ironic English tweets. Negative emotions, in particular, were identified primarily in English *#ironic* self-labelled tweets [23], in different ironic texts in Spanish [22] and Italian ironic tweets [20]. These works show that, among the linguistic strategies that can be used for the activation of irony, some are language-independent, while others seem related to specific languages and cultures. Irony, as a subjective phenomenon, is strongly influenced by individual perception.

The *perspectivist* framework [3] aims at modelling these aspects by incorporating the different points of view represented in the annotations. The new multi-faceted annotation process is then exploited for model training, interpretation, and analysis of the predictions [4]. Perspectivist works on irony are very few. To our knowledge, only two disaggregated datasets for English exist on humour [24] and irony [8]. The first was used as benchmark in the first edition of the LeWiDi (Learning with disagreement) shared task at SemEval 2021; whereas the second was used to build, with a strongly perspectivist approach, demographic-based models to encode annotators' perspectives. Results demonstrated both a variation in the perception of irony based on annotators' social group, and an increase in confidence for perspective-aware models compared to the non-perspectivist ones.

Models	Datasets	iro	non-iro	Annotators	# Annotators	F1-score	Confidence
Fem-persp	FemSet	515	1,450	Self-identified as female	35	.538	.644
Male-persp	MaleSet	536	1,479	Self-identified as male	39	.613	.585
Boomers-persp	BoomersSet	156	283	Of age equal to or above 58	3	.484	.532
GenX-persp	GenXSet	415	1,351	Of age between 42 and 57	22	.483	.612
GenY-persp	GenYSet	577	1,397	Of age between 26 and 41	38	.574	.245
GenZ-persp	GenZSet	322	818	Of age equal to or under 25	10	.601	.352
UK-persp	UKSet	418	955	Of English nationality	15	.533	.630
In-persp	IndiaSet	338	826	Of Indian nationality	15	.432	.708
Ir-persp	IrSet	343	957	Of Irish nationality	15	.521	.340
US-persp	USSet	355	1,004	Of American nationality	14	.461	.583
Au-persp	AuSet	452	916	Of Australian nationality	15	.435	.746

Table 1
Perspective-based datasets, the f1-score and the average of confidence scores obtained by testing the models created on the individual perspectives.

Inspired by their work, and focusing especially on the perception of irony, we propose a linguistic analysis of the predictions of different perspectivist models, which contributes to this emerging framework by examining the most impactful linguistic features for interpreting irony.

3. Dataset and Perspectivist Models

To answer the research questions RQ_1 and RQ_2, we exploit EPIC, the English Perspectivist Irony Corpus released by [8]. This corpus comprises 3,000 pairs of Post-Reply extracted from social media, evenly retrieved from Twitter and Reddit, and was annotated for the irony detection task by crowdsourcing workers with different demographical traits. EPIC was qualitatively examined by [8], that inspected the different demographic-based perspectives encoded in the dataset. They exploited this information to create perspectivist models trained on subsets of data annotated by workers with the same demographical trait. With the aim of examining the perception of irony, we reproduced their perspectivist models and used their predictions for the linguistic analysis.

In more details, following [8] we trained 11 perspective-aware classifiers. Each of these models was trained on data labeled by a specific subset of annotators, who were separated according to their demographic traits as shown in Table 1: gender (female, male), age (boomers, Generation X, Generation Y, Generation Z), and nationality (British, Indian, Irish, American, and Australian). As in [8], we created: i) a unique TEST SET featuring 20% of the instances of EPIC's corpus (246 from Reddit and 307 from Twitter) used for the analyses described in Section 4, ii) and the perspective-specific datasets (see Table 1) by grouping the remaining instance-annotation pairs according to the age, gender, and nationality of their annotators used, in a split 80/20, to train and test the perspectivist models[1].

Each perspective-specific training set was used to fine-tune a pre-trained BERT model [25]. In particular, similar to [8], we finetuned the uncased version of BERT[2] for Sequence Classification, with a binary (ironic and not-ironic) label. Each BERT model was trained by taking as input the representation of the Post-Reply pair. The learning rate was set in a range of 6e-5 and 5e-5, the batch size to 16 and the maximum number of epochs to 10 with an early-stopping strategy.

These models have been tested in perspective-specific test sets, computing the binary label and the confidence score of each model by following [26]'s formula based on the normalized difference between the *logits* of each class, i.e., ironic and not-ironic. The average of the confidence scores over instances and the f1-score of each model are reported in Table 1. As we can notice, the f1-score is fair enough considering the notable unbalance between positive (iro) and negative (non-iro) classes in each dataset.

Once we validated these models, we applied them to the TEST SET (iro: 110, non-iro: 443) obtaining the predictions (and the confidence score of the predictions) of perspectivist models for each instance, like in Table 2[3].

4. Analysis on Perspectives

In this Section, we focus on the analysis of the common and specific patterns that trigger the interpretation of irony of 11 perspective-aware models across the 553 instances of the TEST SET. As commented above, EPIC contains Post-Reply pairs extracted from two sources: Twit-

[1]We note that to label each instance in our perspective-specific datasets, we applied the majority voting strategy to each Post-Reply pair given the annotations of the selected subsets of annotators. We, then, discarded all the entries for which we could not compute a majority vote with the available annotations.
[2]https://huggingface.co/bert-base-uncased
[3]For the sake of clarity, we report the maximum and minimum confidence score only for each instance.

Source	Post	Reply	Perspectivist Models									
			fem	*male*	*boomers*	*genX*	*genY*	*genZ*	*UK*	*In*	*Ir*	*US*	*Au*
Reddit	Other people on social media when they're being trolls. They only do it because 99% of them wouldn't have the nerve to say whatever they're saying to your face.	Saw someone on a friend's FB comments have the nerve to tell her to "check her sources" and link to a meme. The friend has a PhD in the field being discussed.	0	0	0	0	1	0	0	0	0 (.019)	0	0 (.789)
Reddit	Pasta pillows, yes. Pasta cushions even because of the frilly edge. But pasta teabags? No.	Yeah that implies that you dip them in the water and then bin them before drinking your slightly pasta flavoured water.	1	1 (.841)	0 (.003)	0	1	0	0	0	1	0	0
Twitter	Hey atheists, what gives your life meaning if you don't believe in God?	@BeatTheCult Meat,chips, bread and beer....	1	1	1	0	1	1	0 (.781)	0	1 (.044)	1	0
Twitter	Apparently Reece Mogg will be making a statement within the hour. It's not going to be his resignation is it	@YvonneBurdett3 We can only hope! Perhaps we've declared war on Russia or put a man on Mars overnight.	1	1	0	1	0 (.044)	0	1	0	0	0	0 (.781)

Table 2
Predictions in few instances of TEST SET: along with the labels, the minimum and maximum values of the confidence score for each instance.

ter and Reddit. Therefore, we describe two types of analysis: firstly, a source-independent analysis (Section 4.1) and secondly, a source-based analysis (Section 4.2).

The former focused on capturing the linguistic features that trigger the ironic interpretation of a text regardless of its source, exploring the common and diverse features among the predictions of different perspective-based models. The latter aimed at identifying in which source these models tend to predict irony exploring the possible causes, and if there are linguistic patterns specific of a source, looking especially at the use of the strategies and markers identified by [5] in multilingual datasets.

For both analyses, we took into account the predictions of perspectivist models obtained in the TEST SET (Table 2). For each instance, therefore, we have the labels of all the 11 perspectives, and the confidence score of each model computed as described in Section 3. We leveraged the models' knowledge to predict the labels on the TEST SET since – by design –, not all instances of our corpus feature manual annotations covering all demographic traits/perspectives.

4.1. Source-independent Analysis

To observe the commonalities and differences among the interpretation of irony by the various perspectivist models, we extracted a set of linguistic features from the texts of the TEST SET, computed their χ^2 value for each model, and plotted these values in heatmaps[4].

[4]Since we observed that the distribution of the χ^2 values of the features is non-linear, we employed the logarithmic function of PowerTransformer to normalize the data.

To examine the features that are actually discriminative for the detection of irony, we selected for each model only texts from the TEST SET predicted with a very high score of confidence. The threshold used for this selection is unique for each perspectivist model (Table 3), and it was obtained by computing the median of the list of confidence scores resulting from the prediction of positive class (ironic texts) on the specific perspective-based test sets of EPIC (Table 1).

This choice is motivated by one of the findings of [8], who proved that perspectivist models are more confident and precise when predict labels in test sets that encode their perspectives; and depends also on our purpose of examining the perception of irony. We want to be sure that the analysed texts, especially the ones recognized as ironic, have been predicted with a very high confidence by the models.

Models	Threshold	# Texts
Fem-persp	.339	471
Male-persp	.335	439
Boomers-persp	.224	424
GenX-persp	.075	531
GenY-persp	.032	488
GenZ-persp	.091	508
UK-persp	.402	434
In-persp	.254	499
Ir-persp	.031	491
US-persp	.072	531
Au-persp	.179	539

Table 3
Thresholds used to select the most confident predictions for each perspective models.

The selection of the set of features was inspired by existing literature about multilingual and multigenre ironic

texts (Section 2); and include: 1) affective features: the sentiment, emotions, and feelings expressed in the texts (Section 4.1.1); 2) the presence of offensive language (Section 4.1.2); 3) syntactic features (Section 4.1.3). We also performed a lexical analysis (Section 4.1.4).

4.1.1. Affective analysis

We used the EmoLex dictionary [27] to extract emotions and expressed feelings (Figure 1). EmoLex is based on the wheel of emotions theorized by [28], which includes 8 main emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and the primary dyads or feelings (aggressiveness, optimism, love, submission, awe, disapproval, remorse, contempt).

Favored by the design of the wheel of emotions, we computed also the variability of opposite emotions and contrary feelings by means of the standard deviation (σ). The weights of the emotional features are obtained by summing the TF-IDF[5] of words belonging to the specific emotions/feelings. And, we computed the sentiment scores (positive and negative) by using SentiWordNet 3.0 [29] (Figure 2).

As Figure 1 shows, negative emotions and feelings (Example 1) like disgust, contempt, and remorse report **the highest χ^2 values for the majority of the perspectivist models**. Thus, we can confirm the findings of previous analyses in English tweets [23, 14]) where negative emotions were identified primarily in *#ironic* self-labelled tweets. Another common discriminative feature is the contrast between negative emotions and feelings and their positive counterpart (Example 2).

(1) [Post] TLDR: senior positions and management get paid more.
[Reply] And are generally the most useless pricks out there, all talk and no action.

(2) [Post] Fuck carlow they beat me in the feile when I was 13. They all looked like 30 year old men.
[Reply] We have to win a match in football some how.

By looking at the perspective-specific models, we noticed some interesting findings. For instance, when considering the gender dimension, we can notice a higher χ^2 for the Fem-persp model on the presence of negative sentiment and on negative emotions/feelings (fear, sadness, disapproval, and awe) with respect to the Male-persp model (Figures 2 and 1). These values suggest the idea that female annotators tend to recognize irony in texts that express a certain negativity.

Similar finding is noticed in GenY, AU and particularly US-persp models. All these models, indeed, show to be confident in detecting irony when the text is characterized by a negative sentiment, differently from their counterparts (especially GenX, GenZ, IN, IR-persp models). The analysis of emotions brings to light an interesting difference between the IR-persp model and all the 4 models built taking into account the provenance. The IR-persp model shows a marked and higher χ^2 score especially in the presence of emotional contradictions in the texts.

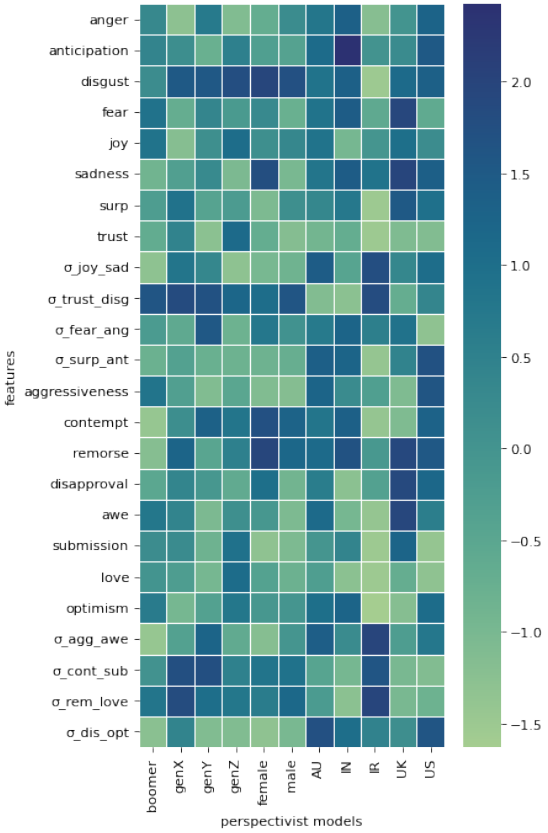

Figure 1: Heatmap visualization for emotion analysis.

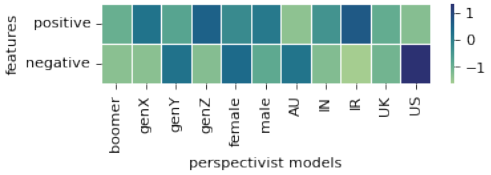

Figure 2: Heatmap visualization for sentiment analysis.

[5]To compute the TF-IDF, we cleaned text from URLs and other not-alphanumeric symbols, tokenized it and removed the stopwords, and finally lemmatized it using the SpaCy large model for English.

4.1.2. Offensive language

The authors of [20] proved that irony, especially in its sarcastic form, can be used to reinforce a negative message. For this reason, the presence of offensive language could be considered a trigger for the ironic interpretation of a text.

To this purpose, we exploited HurtLex, a multilingual lexicon of offensive words. The entries in the lexicon are categorized into 17 types of offences (related to the economic and social spheres, professions, animals, and so on) (Table 4) enclosed in two macro-categories: *conservative* (words with literally offensive sense) and *inclusive* (all the words regardless of the explicitness of the offenses).

Category	Length	Description
PS	254	Ethnic Slurs
RCI	36	Location and Demonyms
PA	167	Profession and Occupation
DDP	496	Physical Disabilities and Diversity
DDF	80	Cognitive Disabilities and Diversity
DMC	657	Moral Behavior and Defect
IS	161	Words Related to Social and Economic advantages
OR	144	Words Related to Plants
AN	775	Words Related to Animals
ASM	303	Words Related to Male Genitalia
ASF	191	Words Related to Female Genitalia
PR	138	Words Related to Prostitution
OM	145	Words Related to Homosexuality
QAS	536	Descriptive Words with Potential Negative Connotations
CDS	2042	Derogatory Words
RE	391	Felonies and Words Related to Crime and Immoral Behavior
SVP	424	Words Related to the Seven Deadly Sins of the Christian Tradition

Table 4
HurtLex categories.

Figure 3 shows that some categories of offensive language report **the highest χ^2 values for the majority of the perspectivist models**. These categories are related in particular to male genitalia, moral behaviors/defects, and, even in its conservative sense, to the category of physical disabilities and diversity.

We can also point out interesting differences **when considering perspective-specific models**. Looking at the gender, we can notice higher values in the Male-persp model when the texts contain words related to crimes/immoral behaviors, professions, and animals, differently from the Fem-persp model.

Observing the dimension of age, instead, the differences are not so marked, except for the offensive words related to female genitalia that appear discriminant for the

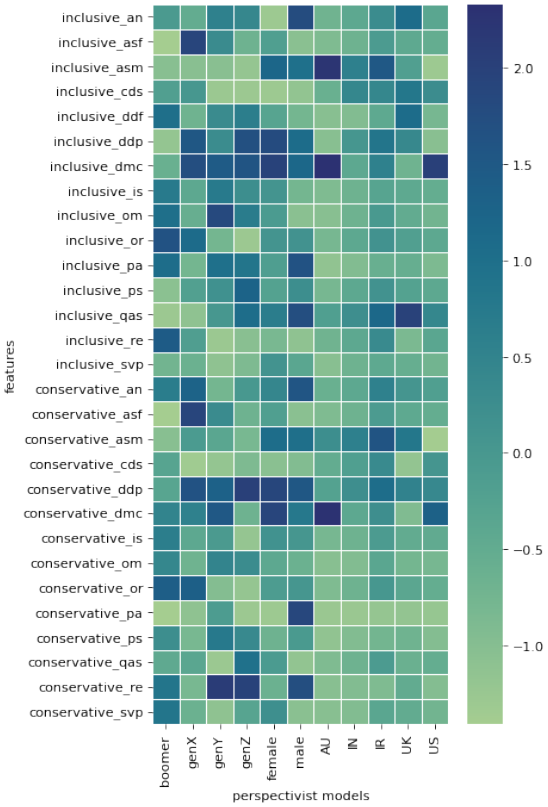

Figure 3: Heatmap visualization of offensive language.

GenX-persp model, and the words related to crimes/immoral behaviours for the youngest generations (i.e., Y and Z). In the dimension of nationality, it is clear that the presence of offensive words related especially to moral behaviours/defects have some impact to the detection of irony for AU and US-persp model. While words related to male genitalia report a higher score only for AU and IR-persp model.

4.1.3. Syntactic features

As shown in previous work [30], syntactic features are proven to be useful to detect ironic language in social media. In particular, we captured syntactic dependencies that could reveal pragmatic information, such as: intensifiers (`intens`), discourse connections (`disc_conn`), adverbial locutions (`adv_loc`), mentions (`mention`) and nominal phrases (and the number of nominal phrases in the tweet) (`nom_phrase` and `num_nom_phrase`). As Figure 4 shows, only the adverbial locutions appear relevant for the majority of models.

However, we noticed that **syntactic features have a higher χ^2 score in a few models**, such as Boomer

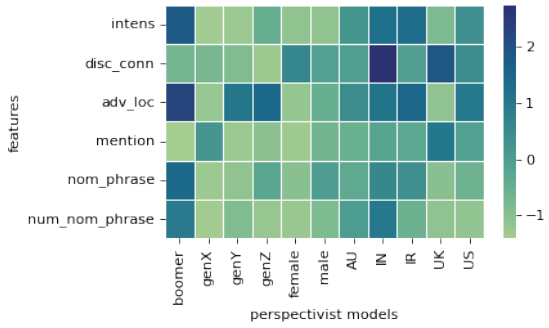

Figure 4: Heatmap visualization of syntactic features.

and IN-persp models. If the former seems to be triggered by different syntactic features (i.e., the presence of intensifiers and nominal utterances), the latter shows to discriminate irony, especially in the presence of discursive connections.

4.1.4. Lexical analysis

To perform a lexical analysis on the TEST SET, we extracted the top 100 unigrams, bigrams and trigrams weighted by their TF-IDF[6] applied separately for each model on texts labelled as ironic. In order to examine the lexical patterns that may influence their choices, we manually analysed both the features that were common to at least 6 models and the ones that occurred in a individual model only.

Focusing on **the n-grams common to at least 5 models**, we individuated a total of 18 features that recur in 5 to 7 models. Ten of them are unigrams frequent across the texts, as *family, think, feel, know*, while the other 8 lexical features are bigrams and trigrams linked to the same 4 texts predicted as ironic by at least 5 models and reported in Table 5.

To highlight whether **some lexical features were model-specific**, we filtered the data by removing all the features that recurred in more than one model of the same dimension (age, gender, and nationality). By manually inspecting these unique features per model, we noticed that for the majority of them, the bigrams and trigrams represented a different combination of the same texts (e.g. *common lannister aside*, *family common lannister*, *lannister aside obsession*, *aside obsession*, *common lannister*, *family common*, *lannister aside*). Boomer-persp and GenY-persp models were the only ones that behaved differently. Their bigrams and trigrams rarely show the systematic repetition of the same lexical items described above, and they both present a higher number of unigrams compared to other models.

[6]We used the TfidfVectorizer from Scikit-learn

Specifically, considering the features associated with the model based on boomers' perspective, there is a high presence of non-English words (as usernames or foreign words, especially from Hindi), and few verbs. In fact, it relies more on nominal n-grams, which in some cases corresponds to the entire text, as in the Examples 3 and 4. This result is further confirmed in the analysis above (Figure 4).

(3) [Post] That's damn shitty of Hugo Boss, what on earth with the chaps in the corner shop and the kebab shop call us now?
[Reply] Ma man

(4) [Post] Election Predictions: Republicans will win the House! Stacey Abrams will lose in Georgia! Any takers?
[Reply] @USER Yo crazy dude

4.2. Source-based Analysis

In this section, we present a quantitative and qualitative analysis of the characteristics of ironic texts in Twitter and Reddit, showing analogies and differences.

4.2.1. Irony on Twitter is more contextual

Observing the predicted texts, we noticed that perspectivist models tend to identify irony more in posts from Reddit (63% of the cases in Table 6) even if the two sources are balanced in the creation process of our corpus.

We hypothesized that this difference was due to the different level of complexity and need for context for instances in the two sources. To measure the characteristics, we computed the length in characters and tokens[7] and lexical richness of the Post-Reply pairs, in terms of type-token-ration (TTR)[8].

We also compute the number of named entities[9] and external elements[10] that could amplify the contextual information in each source (Table 7). We used spaCy and spaCy-udpipe loading the available models for English in particular to extract interjections and the named entities. For the emoticons and emojis, we exploited available lists in the emoji library. While all the other characteristics have been extracted using specific regex.

As expected, posts from Reddit are longer than tweets, but the values of the lexical richness and the number

[7]For computing the length in tokens, the texts have been cleaned and tokenized, removing urls, punctuation, emoji, and emoticons.
[8]TTR is the number of distinct words over the overall words in the text. We took into account tokens and types lists without urls, punctuation, emoji, and emoticons. Here, the texts have been cleaned and tokenized as described in the previous footnote.
[9]The list of named entities considered in this study includes: works of art, organizations, persons, geopolitical entities, locations, events, names of products, date, languages, laws, and nationalities or religious or political groups.
[10]External elements include: hashtags, emoji, emoticons, and urls.

Post	Reply	# Models	Bi/Trigrams
no don't please. i was crushing over since she came. keep that chutiya away	You know there's something else that Trump family has in common with the Lannisters aside from the obsession with gold.	8	*trump family common*
Hey atheists, what gives your life meaning if you don't believe in God?	@BeatTheCult Meat,chips, bread and beer....	7	*meat chip bread*
So BJP-RSS folk need to fear NSG ? Kinda contradictory no ?	Has this guy any shame left. He should be behind bars!	6	*shame leave bar*, *shame leave*
wind power my arse	so....what you think this is false? Or you prefer burning stuff?	8	*prefer burn*, *burn stuff*, *think false prefer*, *prefer burn stuff*

Table 5
Texts predicted as ironic by most of the models, and reporting common relevant bi/trigrams.

# Models	Reddit	Twitter	Tot_Instances
1	53%	47%	271
2	71%	29%	155
3	56%	44%	93
4	77%	23%	48
5	100%	0%	26
6	100%	0%	17
7	50%	50%	7
8	100%	0%	3
Tot	63%	37%	271

Table 6
Distribution of texts classified as ironic by a given number of perspectivist models per source. *# Models* refers to the number of models (i.e., 26 texts have been detected as ironic by at least 5 models, and only 3 texts by 8 models). The columns *Reddit* and *Twitter* refer to the percentage of texts predicted as ironic per source, and *Tot_Instances* refers to the amount of texts detected as ironic: only 271 texts out of 553 have been recognized as ironic by at least one model.

	Reddit		Twitter	
	post	reply	post	reply
length (characters)	207	135	123	85
length (tokens)	38	25	21	14
#named entities	445	244	457	320
#external elements	32	19	170	174
#interjections	35	44	40	57
TTR	0.270	0.306	0.367	0.471

Table 7
Post and replies statistics per source. Numbers correspond to the averages, except for the category of *named entities*, *external elements*, and *interjections*.

of named entities suggest that the content on Twitter is more varied than that from Reddit (Table 7). This is also confirmed by the number of external elements. A similar trend is also observed in the human annotations of the texts of the TEST SET: most annotators recognized more irony in posts from Reddit (27%) than in tweets (14%).

To analyze this trend further, we explored how each model behaves with respect to the source. In general, they identify texts from Reddit as ironic more often than tweets; the only exception is the model trained on the Boomers' perspective, which have classified instances as ironic almost equally for the two sources (52% from Reddit and 48% from Twitter).

4.2.2. Linguistic strategies and markers

We carried out a qualitative analysis of the texts predicted as ironic by at least 5 models, which amounts to a total of 26 texts, 24 from Reddit and 2 from Twitter (Table 6). To these, we added 22 tweets from those identified as ironic by at least 3 models in order to conduct a comparative linguistic analysis of the two sources. For this analysis, we took into account also the irony strategies and markers proposed in the schema of [5] (Section 2).

We found that in both sources, users tend to use similar linguistic strategies to express irony, such as paradox/oxymoron and false assertions, confirming the results presented in [5]; and other interesting features, such as context shift (Example 5) and hyperbole/exaggeration (Example 6).

(5) [Post] How many roads must a man walk down?
[Reply] The only word I know is grunt and I can't spell it.

(6) [Post] Apparently Reece Mogg will be making a statement within the hour. It's not going to be his resignation is it
[Reply] @USER We can only hope! Perhaps we've declared war on Russia or put a man on Mars overnight.

However, some differences are evident. Twitter users often convey contradictions that characterize irony

through unexpected answers (Example 4) and euphemisms (Example 7), while Reddit communities lean towards the use of rhetorical questions (Example 8) and metaphors.

(7) [Post] Lindsey Hoyle spent £7,500 of taxpayers money on a mattress and sheets for his bed in the speakers residence.
[Reply] @USER @USER Very Toriesque

(8) [Post] wind power my arse
[Reply] so....what you think this is false? Or you prefer burning stuff?

From a stylistic point of view, both Reddit and Twitter texts contain question marks, exclamation points, and ellipsis. Full stops are common to the two sources, but they are more frequent in tweets, while Reddit users are more prone to employ swear words.

Tweets also contain nominal utterances more frequently than Reddit posts; this is coherent with the statistics shown in Table 7, which highlight how texts from Reddit are longer and thus include verbal expressions to fulfil complete sentences. In general, in both sources, texts are short and composed of straight answers.

5. Discussion and Conclusion

To the best of our knowledge, this work is the first to approach the analysis of the perceptions of irony in specific segments. Specifically, we base our analysis on the age, gender, and nationality dimension from the EPIC dataset [8]. To examine these patterns in a specific set of texts, we modelled 11 perspectives (self-identified female and male, boomers, generation X, generation Y and generation Z, British, Indian, Irish, American, and Australian), and comparatively analysed the impact of various linguistic features in each of them.

The contribution of this paper is twofold. Firstly, our analysis confirms most of the observations made in the literature about the similar ironic patterns featured in texts of different languages [23, 14, 5]. Secondly, our analysis provides evidence for the different perceptions of irony experienced by people with distinct demographic traits. As a subjective task, irony identification is indeed impacted by experience and background.

Through this analysis exercise, we noticed that the patterns that often trigger ironic interpretation in most perspectivist models are negative emotions (i.e., disgust, contempt, remorse) and contrasting expressions with their counterparts in the wheel of emotions of Plutchik (trust, submission, and love); offensive language (related in particular to male genitalia), moral behaviors or defects, physical disabilities, and diversities also play a role [$\mathbf{RQ}_1$].

In addition, looking at the differences among *perspectives*, we noticed that models trained on female, generation Y, Australian, and American perspectives, often recognize irony when texts convey negative sentiment with respect to their counterparts (respectively, generations X and Z, and Indian and Irish perspectives). Moreover, differently from other models of the provenance dimension, the Irish perspective shows to recognize irony especially in presence of emotional contradictions. In turn, the male perspective model seems more sensitive to irony when the text reports offences related to crimes/immoral behaviors, professions, or animals.

Similar differences are visible in the dimension of age, where texts including female genitalia are considered ironic by Generation X, while the youngest generations (i.e., Y and Z) are more influenced by words related to crimes/immoral behaviors. Finally, only boomers and Indian perspectives are sensible to syntactical patterns, such as intensifiers, nominal utterances, and discursive connectors [$\mathbf{RQ}_2$]. We also noticed that all models detect irony in Reddit posts more often than in tweets.

The findings of these analyses reveal the perception of irony of different segments of people. These observations, therefore, could help to create models for irony detection with different degrees of "subjectivity": models that take into account the most common features to detect irony, or models that target distinct *perspectives*. In both cases, this study provides the *ingredients* to make their decisions explainable. In line with this purpose, we would like, in the future, to enrich these analyses looking also at the topic of the texts, and extend them to different languages, capturing also the understanding of irony in different countries.

Limitations

This work is the first attempt to explore the perception of irony, looking at different perspectives. Given the early stages of this framework, we are aware there are some limitations, which we aim to tackle in subsequent research. In particular, the perspectives are based on a small subset of characteristics (self-identified gender, age, and nationality), and the analysis is conducted using a limited number of data instances (553). To overcome this problem, in the future, we plan to extend these analyses to a larger corpus that includes texts in several languages.

Acknowledgments

The work of S. Frenda, S. Casola and V. Basile was partially funded by the *Multilingual Perspective-Aware NLU* project in partnership with Amazon Alexa. This research was funded through a donation from Amazon.

References

[1] B. Plank, The 'problem' of human label variation: On ground truth in data, modeling and evaluation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Association for Computational Linguistics, 2022. URL: https://arxiv.org/abs/2211.02570. doi:10.48550/ARXIV.2211.02570.

[2] V. Basile, M. Fell, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, A. Uma, We need to consider disagreement in evaluation, in: Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, Association for Computational Linguistics, Online, 2021, pp. 15–21. URL: https://aclanthology.org/2021.bppf-1.3. doi:10.18653/v1/2021.bppf-1.3.

[3] F. Cabitza, , A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, Washington DC, USA, 2023.

[4] G. Abercrombie, V. Basile, S. Tonelli, V. Rieser, A. Uma (Eds.), Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, European Language Resources Association, Marseille, France, 2022. URL: https://aclanthology.org/2022.nlperspectives-1.

[5] J. Karoui, F. Benamara, V. Moriceau, V. Patti, C. Bosco, N. Aussenac-Gilles, Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, pp. 262–272.

[6] A. Joshi, P. Bhattacharyya, M. J. Carman, Investigations in computational sarcasm, Springer Singapore, 2018. URL: https://link.springer.com/book/10.1007/978-981-10-8396-9. doi:https://doi.org/10.1007/978-981-10-8396-9.

[7] 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.

[8] S. Frenda, A. Pedrani, V. Basile, S. M. Lo, A. T. Cignarella, R. Panizzon, C. Marco, B. Scarlini, V. Patti, C. Bosco, D. Bernardi, Epic: Multi-perspective annotation of a corpus of irony, in: ACL 2023, 2023. URL: https://www.amazon.science/publications/epic-multi-perspective-annotation-of-a-corpus-of-irony.

[9] S. Akhtar, V. Basile, V. Patti, Modeling annotator perspective and polarized opinions to improve hate speech detection, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 8 (2020) 151–154. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/7473. doi:10.1609/hcomp.v8i1.7473.

[10] A. T. Cignarella, V. Basile, M. Sanguinetti, C. Bosco, P. Rosso, F. Benamara, Multilingual irony detection with dependency syntax and neural models, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 1346–1358. URL: https://aclanthology.org/2020.coling-main.116. doi:10.18653/v1/2020.coling-main.116.

[11] K. Buschmeier, P. Cimiano, R. Klinger, An impact analysis of features in a classification approach to irony detection in product reviews, in: Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 42–49. URL: https://aclanthology.org/W14-2608.

[12] F. Kunneman, C. Liebrecht, M. Van Mulken, A. Van den Bosch, Signaling sarcasm: From hyperbole to hashtag, Information Processing & Management 51 (2015) 500–509. URL: https://doi.org/10.1016/j.ipm.2014.07.006.

[13] A. Joshi, V. Sharma, P. Bhattacharyya, Harnessing context incongruity for sarcasm detection, 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, Association for Computational Linguistics, Beijing, China, 2015, pp. 757–762. URL: https://aclanthology.org/P15-2124.

[14] D. I. Hernández-Farías, V. Patti, P. Rosso, Irony detection in Twitter: The role of affective content, ACM Transactions on Internet Technology (TOIT) 16 (2016) 1–24. URL: https://doi.org/10.1145/2930663.

[15] S. Zhang, X. Zhang, J. Chan, P. Rosso, Irony detection via sentiment-based transfer learning, Information Processing & Management 56 (2019) 1633–1644. URL: https://www.sciencedirect.com/science/article/pii/S0306457318307428. doi:https://doi.org/10.1016/j.ipm.2019.04.006.

[16] N. Babanejad, H. Davoudi, A. An, M. Papagelis, Affective and contextual embedding for sarcasm detection, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 225–243. URL: https://aclanthology.org/2020.coling-main.20.

[17] 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. URL: https://www.computer.org/csdl/magazine/ex/2013/02/mex2013020055/13rRUxAAT3i.

[18] C. Van Hee, E. Lefever, V. Hoste, SemEval-2018 task 3: Irony detection in English tweets, in: Proceedings of the 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 39–50. URL: https://aclanthology.org/S18-1005. doi:10.18653/v1/S18-1005.

[19] R. Giora, I. Jaffe, I. Becker, O. Fein, Strongly attenuating highly positive concepts. The case of default sarcastic interpretations, Review of Cognitive Linguistics. Published under the auspices of the Spanish Cognitive Linguistics Association 16 (2018) 19–47. URL: https://doi.org/10.1075/rcl.00002.gio.

[20] 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.

[21] A. T. Cignarella, S. Frenda, V. Basile, C. Bosco, V. Patti, P. Rosso, Overview of the EVALITA 2018 task on irony detection in Italian tweets (IronITA), in: Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018), volume 2263, CEUR-WS, 2018, pp. 1–6.

[22] S. Frenda, V. Patti, Computational models for irony detection in three spanish variants, in: CEUR Workshop Proceedings, volume 2421, CEUR-WS, 2019, pp. 297–309.

[23] E. Sulis, D. I. H. Farías, P. Rosso, V. Patti, G. Ruffo, Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not, Knowledge-Based Systems 108 (2016) 132–143. URL: https://doi.org/10.1016/j.knosys.2016.05.035, new Avenues in Knowledge Bases for Natural Language Processing.

[24] E. Simpson, E.-L. Do Dinh, T. Miller, I. Gurevych, Predicting humorousness and metaphor novelty with Gaussian process preference learning, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5716–5728. URL: https://aclanthology.org/P19-1572. doi:10.18653/v1/P19-1572.

[25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[26] A. A. Taha, L. Hennig, P. Knoth, Confidence estimation of classification based on the distribution of the neural network output layer, arXiv preprint arXiv:2210.07745 (2022).

[27] S. M. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon, Computational Intelligence 29 (2013) 436–465. URL: https://doi.org/10.1111/j.1467-8640.2012.00460.x.

[28] R. Plutchik, H. Kellerman, Theories of emotion, volume 1, Academic Press, 1980. URL: https://books.google.it/books?id=TV99AAAAMAAJ.

[29] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf.

[30] A. T. Cignarella, V. Basile, M. Sanguinetti, C. Bosco, P. Rosso, F. Benamara, Multilingual irony detection with dependency syntax and neural models, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 1346–1358. URL: https://aclanthology.org/2020.coling-main.116.