EPIC: Multi-Perspective Annotation of a Corpus of Irony

Simona Frenda[⋆⊙], **Alessandro Pedrani**[⋄], **Valerio Basile**[⋆], **Soda Marem Lo**[⋆], **Alessandra Teresa Cignarella**[⋆⊙], **Raffaella Panizzon**[⋄], **Cristina Marco**[⋄], **Bianca Scarlini**[⋄], **Viviana Patti**[⋆], **Cristina Bosco**[⋆], **Davide Bernardi**[⋄]

[⋆] Computer Science Department, University of Turin, Turin, Italy
[⊙] aequa-tech, Turin, Italy
[⋄] Alexa AI, Amazon, Amazon Development Centre Italy, Turin, Italy

{simona.frenda | valerio.basile | sodamarem.lo | alessandrateresa.cignarella | viviana.patti
cristina.bosco}@unito.it {pedrana | panizzor | marcocri | scarlini | dvdbe}@amazon.it

Abstract

We present EPIC (English Perspectivist Irony Corpus), the first annotated corpus for irony analysis based on the principles of data perspectivism. The corpus contains short conversations from social media in five regional varieties of English, and it is annotated by contributors from five countries corresponding to those varieties. We analyse the resource along the perspectives induced by the diversity of the annotators, in terms of origin, age, and gender, and the relationship between these dimensions, irony, and the topics of conversation. We validate EPIC by creating perspective-aware models that encode the perspectives of annotators grouped according to their demographic characteristics. Firstly, the performance of perspectivist models confirms that different annotators induce very different models. Secondly, in the classification of ironic and non-ironic texts, perspectivist models prove to be generally more confident than the non-perspectivist ones. Furthermore, comparing the performance on a perspective-based test set with those achieved on a gold standard test set, we can observe how perspectivist models tend to detect more precisely the positive class, showing their ability to capture the different perceptions of irony. Thanks to these models, we are moreover able to show interesting insights about the variation in the perception of irony by the different groups of annotators, such as among different generations and nationalities.

1 Introduction

A recent trend in Natural Language Processing (NLP) postulates that the disagreement among annotators in a language resource is a valuable source of knowledge, rather than noise that ought to be minimized or discarded (Plank, 2022; Basile et al., 2021b). Going one step further, the *perspectivist* approach aims at leveraging the disagreement in annotated data in order to model different points of view on the same phenomenon (Basile et al., 2021a). Applied to the study of natural language, this approach is particularly effective when the focus phenomena belong to semantic and pragmatic areas (Abercrombie et al., 2022) such as undesirable language detection, or irony and sarcasm.

Although related, the interpretation of irony involves linguistic patterns, such as the reference to an opposite or secondary meaning, and pragmatic features (Karoui et al., 2017) which could make it possible to recognize the phenomenon for people with different social backgrounds. This differs from the perception of abusive language, proved to be highly affected by different subjectivities (Akhtar et al., 2019). Thus, a fundamental peculiarity of irony is that it tends to be both strongly dependent on the cultural background of the recipients (Joshi et al., 2018; Ortega-Bueno et al., 2019), and, thanks to certain linguistic patterns, it may be understandable regardless of their country of origin.

In this paper, we present EPIC (English Perspectivist Irony Corpus), a corpus of short social media conversations annotated by taking into account the perspective of the annotators. In our view, and according to the perspectivist view, multi-faceted annotation represents an instrument to explore how demographic aspects may influence annotators' opinions, rather than a source of risk of bias. We created EPIC by collecting English messages and their direct replies from public online platforms, and annotated them by crowdsourcing. Crucially, the texts are written in five varieties of English from different countries (Ireland, the United Kingdom, the United States, India and Australia). The annotators, from the same five countries and with different demographic characteristics, expressed their opinion on their perception of irony in texts from all varieties.

We believe that a non-aggregated corpus of irony analysis is a useful resource to train perspective-aware models for irony detection, similarly to the

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 13844–13857
July 9-14, 2023 ©2023 Association for Computational Linguistics

approach of Akhtar et al. (2020) for hate speech modelling. In this direction, we validate the quality of this resource by creating various perspective-aware models for irony detection encoding the perspective of annotators grouped according to their demographic characteristics. These models prove to be more confident in the recognition of irony in comparison with a non-perspectivist model, showing also an interesting increase of the precision in the detection of ironic messages when the various perspectives are represented in the test set. Moreover, the usefulness of EPIC as perspectivist resource is confirmed by the variation in the perception of irony captured through the created perspectivist models.

To sum up, the contributions of this paper are the following: i) a non-aggregated resource for English irony[1]; ii) an analysis of analogies and differences in the annotation on the basis of demographic information about annotators, and correlations between these dimensions and ironic topics; iii) experiments with supervised learning that validate both the quality of the resource and the need for multiple perspectives explicitly encoded in the corpus.

2 Related work

Recent improvements in state-of-the-art language models have shown that the quality of the annotated data required for training automated systems is significantly more important than the amount of data itself (Swayamdipta et al., 2020). For this reason, in NLP, it becomes particularly important to devote special attention to benchmark datasets created within shared tasks and freely available to the research community, as their quality is assessed and improved through multiple uses by researchers. Within the last ~10 years, the amount of irony-annotated resources and the organization of shared tasks regarding figurative language processing (among which, irony and sarcasm) for an increasing amount of different languages has considerably grown. The most resourced language for irony detection is English (Filatova, 2012; Reyes et al., 2012; Van Hee et al., 2016, 2018), but benchmarks have been proposed for other languages, including Spanish (Ortega-Bueno et al., 2019), Italian (Barbieri et al., 2016; Cignarella et al., 2018), Dutch (Van Hee et al., 2016; Maladry et al., 2022), Chinese (Xiang et al., 2020), and Arabic (Alhaidari et al., 2022).

Until 2016, the NLP community has mostly investigated irony as a "general way for describing different kinds of humorous content", (Reyes et al., 2012), as one of the most specific cases of figurative language (Ghosh et al., 2015), or as a "polarity reverser" (Barbieri et al., 2016). Starting from 2017, more specific interest in the phenomenon was deepened, so the community began to study its relationship with sarcasm, hate speech (Van Hee et al., 2018; Cignarella et al., 2018; Frenda et al., 2022), also in different geographical variants of the same language (i.e., Castilian, Mexican, and Cuban variety of Spanish in Ortega-Bueno et al., 2019), and its importance in spreading of stereotypes as well as in author profiling tasks (Ortega-Bueno et al., 2022). As for works on irony that take a perspective approach, we think that the literature on this is not very extensive nowadays; ours is one of the few attempts in this direction. Indeed, after more than a decade of investigation on this subject, it clearly emerged how irony is a highly subjective phenomenon in natural language, for which humans show divergent understanding and interpretation. As with other subjective phenomena, there is therefore an urgent need for the release of datasets with annotator-level labels and socio-demographic information about the annotators (Prabhakaran et al., 2021). A disaggregated dataset about humour in English (Simpson et al., 2019) has been released on the occasion of SemEval 2021 - *Task 12 on Learning with Disagreement* (Uma et al., 2021). However, the currently available lists of disaggregated datasets show that no such kind of dataset exists for irony analysis.[2,3] This paper addresses this issue, since the availability of disaggregated data is a precondition to the study of divergent perspectives on the perception of natural language phenomena (Basile et al., 2021a).

3 Corpus

The corpus we are releasing is called EPIC and is made of $3,000$ short social media text pairs (*Post-Reply*) collected from Twitter ($1,500$) and Reddit

[1] The corpus was exclusively created by the University of Turin, in compliance with the terms and conditions of the data sources. EPIC is made available for research purposes at http://di.unito.it/multilingualperspectivistnlu. Its distribution is governed by the Creative Commons licence. Additionally, the data handling and usage adhere to current regulations, such as the General Data Protection Regulation (GDPR), to ensure the protection of users' rights and privacy.

[2] https://pdai.info/

[3] https://github.com/mainlp/awesome-human-label-variation

(1, 500). Each pair has been annotated by multiple annotators that were asked to provide a binary label (either *Irony* or *not-Irony*) for the *Reply* text given the context provided by *Post*. In the following sections, we describe in detail how we collected the corpus (3.1) and conducted the annotation (3.2).

3.1 Data Collection

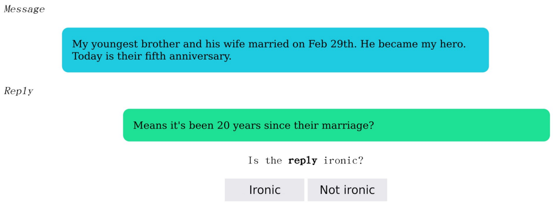
The original data was sourced from two popular social media platforms, namely Reddit[4] and Twitter[5]. The goal was to collect an equal amount of short conversations from social media across the two sources and across five English-speaking countries. To this aim, we collected data from the following subreddits on Reddit, making an assumption about the main origin of their content: r/AskReddit (United States), r/CasualUK (United Kingdom), r/britishproblems (United Kingdom), r/australia (Australia), and r/ireland (Ireland). Furthermore, we collected data from the r/india subreddit, to capture English written by users in India. We downloaded Reddit comments from the archive available in the Pushshift repository[6] selecting the dates between January 2020 and June 2021. We filtered all the comments in the interested subreddits, and saved the (*Post-Reply*) pairs where the *Post* is either a first-level or a second-level comment. Following the collection, we further processed the data by removing all pairs where at least one between *Post* and *Reply* is a deleted or removed comment, and performing a language identification step with the *LangID* Python library[7], retaining only the instances where both *Post* and *Reply* are identified as English.

The data collection from Twitter is designed to yield a result that is as similar as possible to the Reddit section of the dataset. We use the geolocation service provided by the Twitter API to distinguish between English varieties, checking that the country of the (*Post*, *Reply*) pairs corresponds to the target one. We query the Twitter Stream API for tweets in English from each of the five considered countries and retrieve "conversation starting" tweets, i.e., tweets that are neither replies nor quotes. In a second step, we collect the (*Post*, *Reply*) pairs where the *Post* (tweet) is either the conversation starter or a direct reply to it.

After the data collection from Reddit and Twitter, we sampled 600 (*Post*, *Reply*) pairs (300 from Twitter and 300 from Reddit) for each language variety, for a total of 3, 000 instances. Along with the texts, we collected as metadata the subreddit (for the Reddit data), the original post and reply IDs, and the geolocation information (for the Twitter data).

3.2 Annotation

The annotation was conducted through crowdsourcing using a custom-built annotation interface and the service provided by the platform Prolific[8]. The annotation interface is designed to draw instances from a relational database, selecting a random instance which i) has not been already annotated by the current user, and ii) does not show more than a predetermined number of annotations. Each instance to annotate is composed of a *Post* and a *Reply*, which are shown on screen in a way that emulates message chats. When presented with an instance, the user is simply asked to select whether the *Reply* is ironic or not[9], by clicking on one of two buttons — see Figure 1 for a screenshot of the interface seen by the annotators. The custom software is integrated with the API provided by Prolific, exchanging only an anonymized user ID, and redirecting the user to the payment page once the task is complete.



Message

My youngest brother and his wife married on Feb 29th. He became my hero. Today is their fifth anniversary.

Reply

Means it's been 20 years since their marriage?

Is the **reply** ironic?

Figure 1: Screenshot of the annotation interface.

For the annotation of EPIC, we decided to hire a total of 76 annotators, 16 from the United Kingdom, and 15 from each of the remaining interested countries[10]. Each instance is annotated by five different annotators, and each annotator completed 200 annotations.

We selected the annotators so that they are native speakers of English, and have a task completion rate on other Prolific tasks of 99%, as a filter for quality. We asked the crowdsourcing platform to

[4] https://reddit.com/

[5] https://twitter.com/

[6] https://redditsearch.io/

[7] https://github.com/saffsd/langid.py

[8] https://prolific.co

[9] The instructions for the annotation process are shown in Appendix A(1).

[10] The platform rejected one annotator from the UK based on a time limit. However, since their annotation was completed, we included it in the dataset (and paid the annotator).

provide balanced sets of annotators with respect to their gender, but left the other filters open, in order to capture wider demographics.

We did however force a balance across the country of residence of the annotators.[11] This choice concerns the design of the resource, and it is fundamental for the aim of considering multiple perspectives on the perception of irony. Annotators had to annotate instances from all five varieties of English, not just the one they speak as native speakers, and we designed the software to balance the countries of the annotators when assigning new instances to them.

To further guarantee the reliability of the annotations, we included attention-check questions. Together with the task completion rate, they have been used to ensure the quality of the corpus while keeping the data disaggregated, coherently with the *perspectivist* approach. For each new question, the annotators have a 1% probability of receiving an attention check instead of an actual instance of the dataset to annotate. The attention-check questions have the form "please reply [yes/no] to this question". We chose a threshold of 50% correct answers in order to consider the annotator valid. Among the 76 annotators, just two of them failed the test, resulting in a total of 74 annotators.

4 Statistical analysis

EPIC contains $3,000$ unique annotated instances (*Post*, *Reply*) collected and annotated as described in sections 3.1 and 3.2. In this section, we provide high-level statistics about annotators and annotations and explore annotations at a deeper level. Similarly to Prabhakaran et al. (2021), we prove that aggregation by majority voting would introduce representational biases of individual and group viewpoints. In addition, we show how annotators' perceptions differ depending on the topic for which irony is being labelled.

Annotators' Summary Statistics We recorded basic demographic information for the pool of 74 retained annotators. In particular, we observed: *Gender* (39 Males, 35 Females), *Age Group* (38 Gen-Y, 22 Gen-X, 11 Gen-Z and 3 Baby Boomer, 1 Null[12]), *Nationality* (15 United Kingdom, 15 India, 15 Ireland, 15 Australia, 14 United States), *Ethnicity* (47 White, 18 Asian, 3 Black, 6 Other or Null), *Student Status* (46 No, 13 Yes, 15 Null) and *Employment Status* (24 Full-Time, 11 Part-Time, 11 Unemployed, 4 Not in paid work, 24 Other or Null). We recognize how 74 is not a huge number for annotators. However, it is sufficient to observe statistically significant differences among groups (see section 4). In addition, perspectives considered later (in section 5) are modelled along axes that are orthogonal to each other, leading to small but sizeable enough subgroups. For instance, 'gender' and 'nationality' (almost perfectly balanced), together with 'age' (unbalanced, but with only the *boomer* class being underrepresented).

Annotations Summary Statistics Overall, we recorded $14,172$ annotations. Each instance has on average 4.72 annotations, with the median being 5. The first remarkable fact is the disagreement among annotators. More than 66% $(2,010)$ of the instances have at least one annotator disagreeing with the others, and 30% of texts with more than four annotations (868 our of $2,784$) have at least two annotators voting both *Irony* and two voting *not-Irony*. Calculating the majority label for each instance as the label that half or more annotators who annotated that instance agreed on, results in 649 instances being labelled as *Irony*, $2,118$ as *not-Irony* (233 remaining are ties).

Majority vote introduces Bias Prabhakaran et al. (2021), showed that the majority vote underrepresents or ignores the perspectives of a sizeable number of annotators, at least on datasets for 3 tasks on which they focused: Hate-speech, Sentiment and Emotions recognition. We proved that their findings hold true for irony on EPIC. To this end, we compute Cohen's κ agreement score for each annotator by comparing the list of labels provided by the individual, with the list of majority-vote labels on the subset of instances for which the annotator provided a label. Figure 2 represents the histogram and Kernel Density Estimation of annotators' Cohen's κ agreement score with majority votes. While a certain level of disagreement is expected, and can be attributed to noise (e.g. annotators' errors), the overall assumption of a majority vote aggregation is that it captures the perspective of the *average annotator* within a pool.

However, we observed that such a majority voting scheme will not uniformly represent all groups

[11] For contributors from India we used 'nationality' instead of 'residence' since no annotators residing in India were available on Prolific.

[12] One of the annotators did not share this information. This annotator was included in the statistical analyses, except for the one related to 'age'.

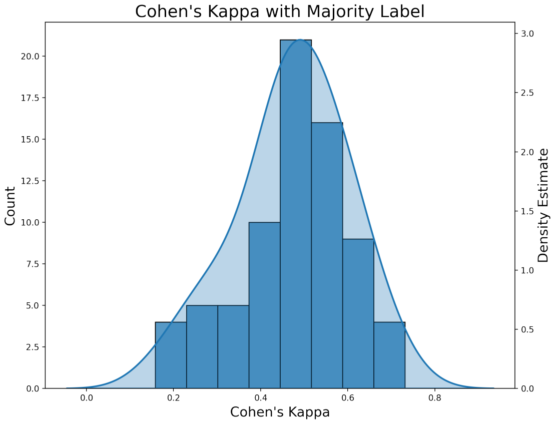


Figure 2: Histogram and kernel density estimation of the distribution of Cohen's κ agreement score of each annotator versus the majority labels.

in the pool. Violin plots in Figure 3 show an estimate of the distribution of the Cohen's κ score with majority votes for annotators across different classes: *Gender*, *Age Group*, *Nationality*, *Ethnicity*, *Student Status* and *Employment Status*. These plots suggest that there is a remarkable qualitative difference in how the groups are represented by the majority votes. For instance, even though *Males* and *Females* almost have the same average agreement (0.466 vs 0.478), there is an evident difference in variance, with *Females*' scores being more concentrated. We also observed that the perspective of annotators self-identifying as *Asians* (average 0.414) is way less represented by the majority voting than the perspective of annotators self-identifying as *White* (average 0.493). A Welch's t-test Welch (1947) suggests a significant difference between the two groups (p-values are 0.026). Similarly, annotators whose nationality is *India* (average 0.413) are way less represented by majority labels than annotators from *Ireland* (average 0.500), even though in this case the statistical test report a p-value on the boundary of the conventional 0.05 threshold (precisely 0.062) suggesting a slightly higher chance of type I error in considering the two groups as different.

Agreement depends on the Topic In order to verify if agreement, and therefore irony perception, also depends on the topic of the corpus being annotated, we classified instances into topics. Since our primary goal here is interpretability, we adopted a simpler but solid approach to topic modelling. First, we selected the first level of the taxonomy of topics of media news as defined by the *International Press Telecommunications Council*[13]. This resulted in a pool of 18 topics: *arts, emergency, economy, education, environment, health, human interest, justice, labour, lifestyle, politics, religion, science, society, sport, technology, war, weather*. Then, we followed the approach described by Yin et al. (2019) and used a pre-trained Natural Language Inference model as a zero-shot sequence classifier to classify our instances into the above list of topics. In particular, we used `facebook/bart-large-mnli`, that is the fine-tuned version of `bart-large` Lewis et al. (2019) trained on the MultiNLI dataset Williams et al. (2018). This is publicly available in the *Hugging Face*[14] repository. We then associated to each text the top three topics proposed by the model with a score > 0.5[15]. Figure 4 shows the resulting distribution of topics, where human interest, environment, and lifestyle are the more frequent ones.

For each instance i, we considered the set of annotators A providing a label for i and computed a measure of agreement a between them on instance i as:

$$a_i = 1 - \frac{\chi_i^2}{|A|}$$

where χ_i^2 is the value of the χ^2 statistics to test if labels assigned are from a uniform distribution. This is inspired by Akhtar et al. (2019). Note how a_i will be 1 if annotators are in perfect disagreement (50% annotated *Irony* and 50% annotated *not-Irony*) while will be 0 if annotators are in perfect agreement (all of them annotated *Irony* or all of them annotated *not-Irony*). We do not use Cohen's κ agreement score to measure agreement, since this is a property of each annotator. Rather, we compute the agreement of multiple annotators on the same instance (and topic). Therefore, we proceed by computing the average polarization by topic — the result is shown in Figure 5.

Some topics such as *labour* ($p = 0.614$), *science* ($p = 0.600$), *lifestyle* ($p = 0.575$), *emergency* and ($p = 0.572$) *politics* ($p = 0.571$) exhibit a remarkably higher polarization than others, such *health* ($p = 0.478$) and *arts* ($p = 0.459$). These results show the need to release perspectivist datasets.

[13] https://iptc.org/

[14] https://huggingface.co/

[15] Some examples are reported in Appendix A(2).

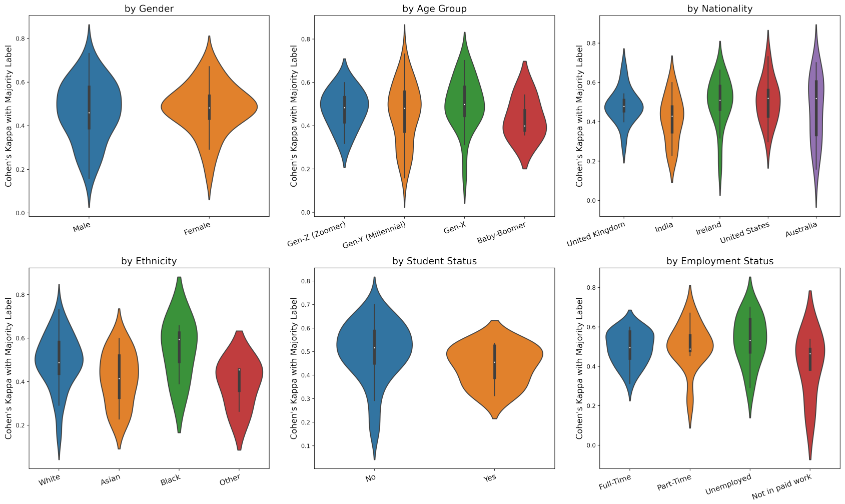


Figure 3: Violin plots showing the distribution of the Cohen's κ agreement score across different dimensions. We can appreciate how in multiple cases the distributions are much different, revealing that majority voting aggregation would systematically penalize certain groups of annotators. A global agreement measure is omitted as not meaningful in the perspectivist framework.

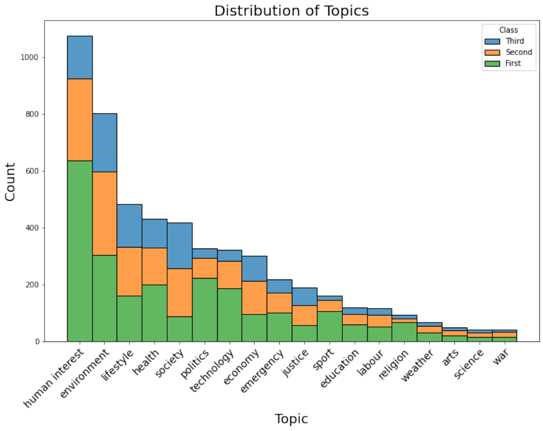


Figure 4: Distribution of topics per instances in the dataset.

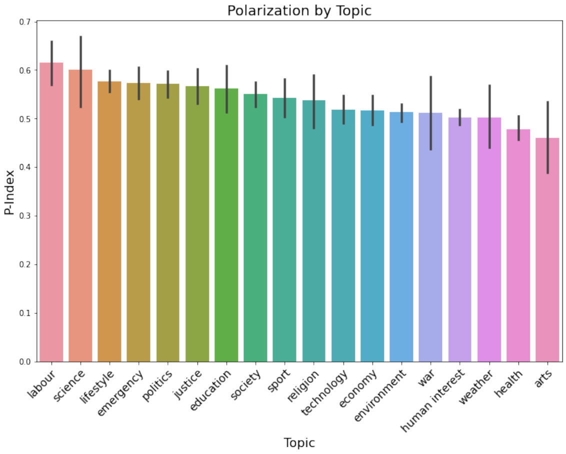


Figure 5: Polarization across Topics.

5 Perspective-aware modelling results

In this section, we describe computational experiments to detect irony using the EPIC dataset. As described above, this dataset has been annotated by different annotators coming from five English-speaking countries and with different demographic characteristics. Using the available information, we designed several classifiers that take into account the subjectivity of various groups of annotators divided according to their demographic characteristics.

Indeed, the EPIC dataset offers the opportunity to explore perspectivist approaches for irony detection, exploiting the information available about annotators. In these experiments, we want to understand the importance of a perspectivist approach for irony detection compared to a standard non-perspectivist approach, whose training and testing are based on a gold standard dataset. In particular, we want to answer the following questions: **(1)** What is the difference, especially in terms of confidence, between perspectivist and non-perspectivist models? **(2)** Along which dimension can we observe the highest variation in the perception of irony?

The first step was the creation of specific datasets to train and test the perspective-aware models, grouping the annotated texts on the basis of age, gender, and provenance of annotators as shown in Table 1. To get a pair text-label in our datasets, we applied the majority voting strategy to each slice and discarded the instances for which we cannot compute a majority vote with the available annotations. A gold standard dataset (called here Gold-Set) was also produced to create a non-perspectivist model. In this dataset, the pair text-label was designed employing a majority voting among all the decisions collected by annotators regardless of their characteristics.

Dataset	# Instances	Annotators
GoldSet	2,767	All the annotators, only instances with 5 or more annotations with fully aggregated labels.
FemSet	1,952	Self-identified as female.
MaleSet	2,023	Self-identified as male.
BoomersSet	441	Older than 58.
GenXSet	1,757	Older than 42 and younger than 57.
GenYSet	1,964	Older than 26 and younger than 41.
GenZSet	1,124	Younger than 25.
UKSet	1,365	With English nationality.
IndiaSet	1,175	With Indian nationality.
IrSet	1,296	With Irish nationality.
USSet	1,352	With American nationality.
AuSet	1,377	With Australian nationality.

Table 1: Datasets extracted from EPIC.

Our experiments consist of a fine-tuning of the pre-trained BERT (Devlin et al., 2019) for English language on each of these datasets to create different perspective-based models to detect irony in English tweets and posts from Reddit. For the training phase of each model (perspectivist and not), we selected a training and validation set[16] corresponding to the 80% of the dataset. For the testing phase, we selected a GOLD TEST SET from the GoldSet of 553 instances corresponding to 20% of the entire GoldSet and a PERSPECTIVE-BASED TEST SET from each subjective set of data (the 20% of each dataset). According to this, all the perspective-based datasets in Table 1 have been created excluding the instances of the GOLD TEST SET. The training, validation, and test set have been balanced on the basis of the source: Twitter and Reddit. The employed language model, the description of the input, the hyperparameters' values and the functions used in these experiments are presented in the Appendix A(3). This experimental setting includes the application of early-stopping strategy to avoid the overfitting in the training phase of the models.

To answer the first question, we compare the performance of perspective-aware models on both the PERSPECTIVE-BASED TEST SET and GOLD TEST SET. The performance on the latter are further compared with the model obtained fine-tuning BERT on the training set of GoldSet (the *non-perspectivist model*). For the evaluation, we report the *F1-score* measure, but we focus, especially, on the average (*avg*) and standard deviation (*std*) of the confidence scores of all the predictions in order to gauge the degree of certainty/uncertainty of the models on both test sets. In Table 2, we also reported the percentage of variation of model confidence in terms of Δ. The confidence score of each prediction is computed using the formula proposed by Taha et al. based on the normalized difference between the *logits* obtained for each class (ironic and not-ironic). The logits have been rescaled by applying the softmax function.

Looking at Table 2, firstly, we can notice that Male-persp model performs better on the GOLD TEST SET, even if: the distribution of annotations on the basis of genre (between male and female annotators) has been required to be balanced in Prolific platform (see Section 3.2); the amount of annotated data in FemSet and MaleSet is similar (see Table 1); and even if the IAA among female annotators show to be more consistent than male annotators (see Figure 3). Along with Male-persp model, also the GenY-persp reports a F1-score greater than 0.60. These two perspectives seem

[16] The validation set was employed to stop the fine-tuning of the model in the frame of an early-stopping strategy.

model	GOLD TEST SET			PERSPECTIVE-BASED TEST SET				
	F1-score	Confidence		F1-score	Confidence		Δ% Confidence	
		std	*avg*		*std*	*avg*	*std*	*avg*
non-perspectivist	0.681	0.301	0.509	–	–	–	–	–
Fem-persp	0.590	0.239	0.621	0.538	0.234	0.644	-2.09↓	3.70↑
Male-persp	0.620	0.274	0.582	0.613	0.267	0.585	-2.55↓	0.52↑
Boomers-persp	0.539	0.290	0.502	0.484	0.303	0.532	4.48	5.98↑
GenX-persp	0.516	0.269	0.603	0.483	0.261	0.612	-2.97↓	1.49↑
GenY-persp	0.611	0.265	0.255	0.574	0.259	0.245	-2.26↓	-3.92
GenZ-persp	0.574	0.234	0.367	0.601	0.240	0.352	2.56	-4.09
Au-persp	0.497	0.173	0.748	0.435	0.165	0.746	-4.62↓	-0.27
US-persp	0.516	0.259	0.580	0.461	0.262	0.583	1.16	0.52↑
Ir-persp	0.535	0.273	0.319	0.521	0.293	0.340	7.33	6.58↑
In-persp	0.466	0.232	0.666	0.432	0.210	0.708	-9.48↓	6.31↑
UK-persp	0.507	0.255	0.612	0.533	0.251	0.630	-1.57↓	2.94↑

Table 2: Classification performance and confidence of perspective-aware models *vs.* non-perspectivist model.

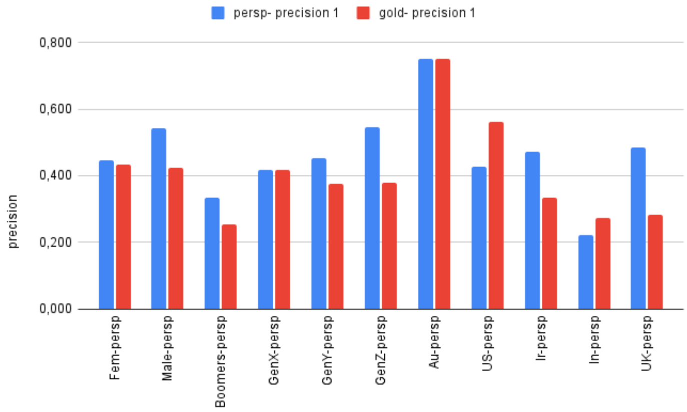


Figure 6: Precision performance on positive class of perspective-aware models.

to be present more than others in the GOLD TEST SET. However, it is interesting to notice that none of the perspectivist models perform better than the non-perspectivist model on the GOLD TEST SET because the *gold* labels are not representative of each specific perspective. Another interesting point is the high variability on the GOLD TEST SET of the performance of the models built taking into account decisions of annotators with different traits. That means that different annotators induce very different models.

Secondly, two important trends are visible in the GOLD TEST SET column: the standard deviation and the average of confidence scores appear, respectively, lowering (↓) and increasing (↑) in the performance of perspective-aware models respect to the performance of the non-perspectivist model. That means perspective-aware models tend to take a decision with less uncertainty than standard non-perspectivist models. A similar result was expected observing the percentage of Δ between the *avg* and *std* of confidence scores, where we can show that perspective-aware models are inclined to be respectively more confident and consistent when they are tested on a test set representative of their perspective. To examine in depth this result, we look also at the performance on positive class (*ironic texts*) of perspective-aware models, reporting in Figure 6 the *precision* scores of ironic class obtained on the PERSPECTIVE-BASED TEST SET (blue bars) and on the GOLD TEST SET (red bars). In this figure, the blue bars tend to be higher than the red ones in the majority of the cases, suggesting that the different perceptions of irony can be well recognized by perspective-aware models. We observe an increase in Δ in a range from 3% with the Fem-persp model to 72% with the UK-persp model.

To answer the second question, we compared the different and similar predictions obtained from perspective-aware models of the same category (gender, age, and country). In the previous sections, we looked at the difference in IAA among different groups of the same demographic category. Now, we focus especially on the variation of their perception of irony captured by perspective-aware models. To this purpose, we computed the accuracy measure among the predictions obtained with the various perspectivist models on the GOLD TEST SET.

	male
fem	.85

	genX	genY	genZ
boomers	.73	.71	.81
genX	–	.80	.87
genY	–	–	.79

Table 3: Variation among perspectives on ‘gender’ (left) and perspectives on ‘age’ (right).

	US	Ir	In	UK
Au	.96	.91	.97	.93
US	–	.91	.95	.92
Ir	–	–	.89	.88
In	–	–	–	.89

Table 4: Variation among perspectives on ‘nationality’.

Looking at Tables 3 and 4 reporting the variation among perspectives on the demographic categories, we can observe some differences of perception of irony (in a range from 3% to 29%), especially on ‘gender’ and ‘age’. For instance, contiguous generations seem to perceive irony in different way (i.e., boomers vs. genX, genX vs. genY, genY vs. genZ), although boomers vs. genY results in the highest variation. Interestingly, looking at the countries, the highest variation, even if less strong than for ‘age’, is reported between the predictions of the models trained on annotators’ decisions coming from United Kingdom and Ireland.

All these findings prove the necessity to take into account the different perspectives of people to create more confident and representative models, even in a difficult task such as the recognition of irony.

6 Conclusion

In this paper, we presented EPIC, a corpus of short social media conversations from five English varieties (Australian, British, Indian, Irish, American) collected from Twitter and Reddit and annotated with a binary label, *Irony* or *not-Irony*, by speakers from the five countries. We performed statistical analyses resulting in two key takeaways. The first is that aggregating the dataset with a majority voting scheme would introduce biases, thus hiding the perspective of some groups of annotators (e.g., those identifying as Asian). This confirms the hypothesis that the perception of Irony is dependent on the cultural background of the recipient. The second is that polarization among annotators depends on the topic. This means that though it is true that cultural background influences the perception of ironic content, there exist topics (such as Arts and Health) on which the influence is less evident than on others (such as Labour, Lifestyle or Politics). Moreover, we performed predictive experiments creating perspective-aware models for irony detection, that show how different annotators induce very different models, and how these perspectivist models, trained on subsets of the annotation coming from identifiable perspectives, are more confident at prediction time. Finally, looking at the detection of irony, we believe that the best approach is based on assembling perspective-aware models plus perspective-based explanations. This is beyond the scope of the current work, which wants to present a solid basis on which to build such models.

We plan to continue our research in two main directions. Firstly, we intend to expand the dataset beyond English (i.e., Spanish, German, French, Italian, Arabic, and others) in order to create the first multilingual perspectivist dataset for irony detection. Secondly, we will employ EPIC as the basis for more advanced perspective-aware models and as a perspectivist benchmark for irony detection.

Limitations

While this work represents the first effort towards a perspectivist language resource for irony detection, it has to be noticed that the resource is monolingual (English). Moreover, while we tried to maintain a fair balance in terms of demographic profile of the annotators, we limited the resource to five varieties of English tied to five countries, while leaving out other potential locations (e.g., New Zealand or Nigeria) or even more nuanced distinctions among language varieties. About the self-identified gender dimension, we are aware of the wider spectrum of genders. However, this information is provided by the annotators only in a binary form. Another potential limitation is that, in the spirit of constructing a perspectivist corpus, we fully trusted the contributors. While the chosen crowdsourcing platform (Prolific) is known for a high quality standard obtained e.g. by vetting its contributors, and we added a layer of checks through attention test questions, random noise in the annotation may still be present and undetected.

While this paper mainly presents a new language resource, we also included the results of several analyses and validation experiments. In this direction, a number of dimensions are still unexplored, along which the data could be analysed. For instance, the genre difference between the sources of the data (Reddit and Twitter) and the distribution of different varieties of English were not yet explored.

Ethics Statement

The research presented in this paper relies on the labour of numerous contributors who annotated the dataset. We recruited and rewarded our contributors through Prolific, a crowdsourcing platform we selected specifically for its attention to fair and ethic treatment of crowdworkers. The contributors were paid on average an hourly wage of 12.66 GBP (about 14.95 USD). Additionally, fixed bonus payments were provided for contributors who abandoned the task but still provided valuable feedback.

The data perspectivist approach in general, and this work in particular, aims at "giving voice to the few who hold a minority view" (Basile et al., 2021a). Applied to the creation of a language resource, this principle leads to resources (and therefore models) where bias is a controlled factor rather than undesirable criticality.

Acknowledgements

The work of S. Frenda and V. Basile, C. Bosco, A.T. Cignarella and V. Patti was partially funded by the *Multilingual Perspective-Aware NLU* project in partnership with Amazon Alexa. The work of A.T. Cignarella, V. Patti, and C. Bosco was partially funded by the International project *STERHEOTYPES - Studying European Racial Hoaxes and sterEOTYPES*, funded by the Compagnia di San Paolo and VolksWagen Stiftung under the 'Challenges for Europe' Call for Projects (CUP: B99C20000640007). This research was funded through a donation from Amazon.

References

Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors. 2022. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, Marseille, France.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *AI*IA 2019 – Advances in Artificial Intelligence*, pages 588–603, Cham. Springer International Publishing.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):151–154.

Linah Alhaidari, Khaled Alyoubi, and Fahd Alotaibi. 2022. Detecting irony in arabic microblogs using deep convolutional neural networks. *International Journal of Advanced Computer Science and Applications*, 13(1).

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *Proceedings of 3rd Italian Conference on Computational Linguistics (CLiC-it 2016) & 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*. CEUR-WS.org.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021a. Toward a perspectivist turn in ground truthing for predictive computing. *CoRR*, abs/2109.04270.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021b. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2018. Overview of the EVALITA 2018 Task on Irony Detection in Italian Tweets (IronITA). In *Proceedings of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*. CEUR-WS.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Elena Filatova. 2012. Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 392–398. European Language Resources Association.

Simona Frenda, Alessandra Teresa Cignarella, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. Semeval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. ACL.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2018. *Investigations in computational sarcasm*. Springer Singapore.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. 2017. 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*, pages 262–272, Valencia, Spain. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Aaron Maladry, Els Lefever, Cynthia Van Hee, and Veronique Hoste. 2022. Irony Detection for Dutch: a Venture into the Implicit. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 172–181.

Reynier Ortega-Bueno, Berta Chulvi, Francisco Rangel, Paolo Rosso, and Elisabetta Fersini. 2022. Profiling irony and stereotype spreaders on twitter (irostereo). In *CLEF 2022 Working Notes*, volume 3180. CEUR-WS.

Reynier Ortega-Bueno, Francisco Rangel, D Hernández Farıas, Paolo Rosso, Manuel Montes-y Gómez, and José E Medina Pagola. 2019. 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, pages 229–256.

Barbara Plank. 2022. 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.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From Humor Recognition to Irony Detection: The Figurative Language of Social Media. *Data & Knowledge Engineering*, 74:1–12.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5716–5728, Florence, Italy. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Abdel Aziz Taha, Leonhard Hennig, and Petr Knoth. 2022. Confidence estimation of classification based on the distribution of the neural network output layer. *arXiv preprint arXiv:2210.07745*.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2016. Exploring the realization of irony in Twitter data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1794–1799. European Language Resources Association.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation (SemEval 2018)*, pages 39–50. ACL.

Bernard Lewis Welch. 1947. The generalization of 'student's' probem when several different population variances are involved. *Biometrika*, 34(1-2):28–35.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Rong Xiang, Xuefeng Gao, Yunfei Long, Anran Li, Emmanuele Chersoni, Qin Lu, and Chu-Ren Huang. 2020. Ciron: a New Benchmark Dataset for Chinese Irony Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. ELRA.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

A Appendix

1. Instructions for the annotation process

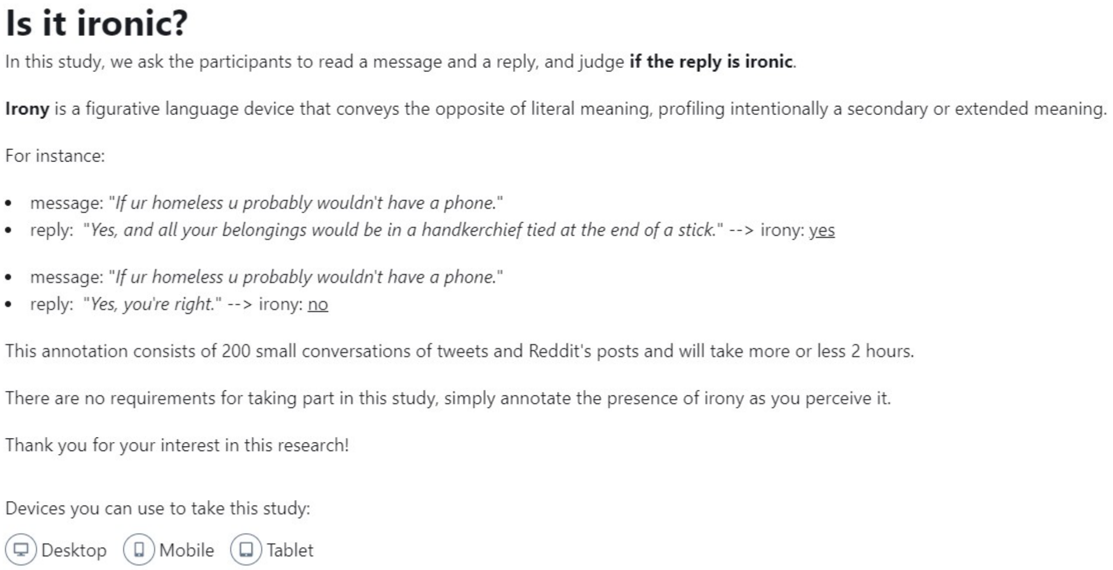
Figure 7 shows the instructions as seen by the annotators in Prolific before they choose to undertake the task.

2. Examples of topic classification

Table 5 reports some example of the topic classification described in Section 4.

3. Language Model Parameters

Table 6 shows the values of the hyperparameters used in the experiments presented in Section 5.



Is it ironic?

In this study, we ask the participants to read a message and a reply, and judge **if the reply is ironic**.

Irony is a figurative language device that conveys the opposite of literal meaning, profiling intentionally a secondary or extended meaning.

For instance:

- message: "*If ur homeless u probably wouldn't have a phone.*"
- reply: "*Yes, and all your belongings would be in a handkerchief tied at the end of a stick.*" --> irony: yes

- message: "*If ur homeless u probably wouldn't have a phone.*"
- reply: "*Yes, you're right.*" --> irony: no

This annotation consists of 200 small conversations of tweets and Reddit's posts and will take more or less 2 hours.

There are no requirements for taking part in this study, simply annotate the presence of irony as you perceive it.

Thank you for your interest in this research!

Devices you can use to take this study:

Figure 7: Instructions for the annotators in Prolific.

Post Text	**Reply Text**	**Topic 1**	**Topic 2**	**Topic 3**
The NFL is rigged. I mean, there's too much money on the line per game for there not to be someone wanting to fix it. [..] Super Bowls are blowouts or close games based on the highest payers' time slots in the game. [...]	All valid points.	sports	N/A	N/A
Probably BoTW and Minecraft	Yup	technology	N/A	N/A
The Jews control Israel.	I mean, you're not wrong, but...	religion	politcs	N/A
Travellers have been lobbying for a national health strategy, mental health strategy for over a decade our State and its organs failed us. Now look where we are our children dying by suicide at a shocking rate.	those poor children, it's time for some intervention	health	emergency	human interest

Table 5: A sample with 4 examples of (*Post*, *Reply*) instances in the dataset and their classification with our topic extraction approach. Though not perfect, the resulting classification is satisfactory and being highly interpretable is adequate for our needs.

parameter	**value**
model	the uncased version of BERT (https://huggingface.co/bert-base-uncased) for Sequence Classification, predicting 2 labels (ironic and not-ironic) for each text.
input	the pair Post-Reply, reproducing the input of the annotation phase as shown in Figure 1 and giving contextual information to the system.
max sequence length	100
learning rate	[6e-5, 5e-5]
batch size	16
maximum number of epochs	10
optimizer	AdamW
scheduler	the *cosine scheduler* without warmup (https://huggingface.co/transformers/main_classes/optimizer_schedules.html) to define dynamic learning rates during the training phase.
early stopping	a custom early stopping function to avoid the overtraining of the neural network, looking at the values of the loss obtained on the validation set with a *patience* of 3 epochs.
seed	a constant *seed* to make the results reproducible.
loss	the default loss function defined for Sequence Classification by *transformers* library.

Table 6: Language model, parameters' values and functions used for the fine-tuning process.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section titled "Limitations" after Section 6 (Conclusion).

☒ A2. Did you discuss any potential risks of your work?
We do not foresee any potential risks involved in the use of the resource.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1 (Introduction).

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We created a language resource described in Section 3 (Corpus). We used pretrained language models in Sections 4 and 5.

☑ B1. Did you cite the creators of artifacts you used?
Section 4 and 5.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 1 (Introduction)

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We made a standard use of scientific artifacts employed in this paper, in accordance with their terms of use.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Our resource contains not-anonymized social media data collected from public forums. However, we will follow the General Data Privacy Regulation as indicated in Section 1.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sections 3 and 4.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Sections 3, 4 and 5.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Sections 4 and 5.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We reported the hyperparameters used in the experiments in the Appendix A(3). We only ran fine-tuning experiments with negligible computational costs (a few hours on a single GPU).

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sections 4, 5 and Appendix A(3).

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3.2

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3 and Appendix A(1).

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 3 and Appendix A(1).

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Following the European regulations we do not consider necessary the approval by an ethics review board at the time of the submission. We received the approval of IP and Legal review board.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3 and 4.