Human and System Perspectives on the Expression of Irony: an Analysis of Likelihood Labels and Rationales

Aaron Maladry[♢], **Alessandra Teresa Cignarella**[♣], **Cynthia Van Hee**[♢], **Els Lefever**[♢] and **Veronique Hoste**[♢]
♢ LT3, Ghent University, Belgium
♣ aequa-tech, Turin, Italy
aaron.maladry@ugent.be
alessandrateresa.cignarella@aequa-tech.com
cynthia.vanhee@ugent.be
els.lefever@ugent.be
veronique.hoste@ugent.be

Abstract

In this paper, we examine the recognition of irony by both humans and automatic systems. We achieve this by enhancing the annotations of an English benchmark data set for irony detection. This enhancement involves a layer of human-annotated irony likelihood using a 7-point Likert scale that combines binary annotation with a confidence measure. Additionally, the annotators indicated the trigger words that led them to perceive the text as ironic, which leveraged necessary theoretical insights into the definition of irony and its various forms. By comparing these trigger word spans across annotators, we determine the extent to which humans agree on the source of irony in a text. Finally, we compare the human-annotated spans with sub-token importance attributions for fine-tuned transformers using Layer Integrated Gradients, a state-of-the-art interpretability metric. Our results indicate that our model achieves better performance on tweets that were annotated with high confidence and high agreement. Although automatic systems can identify trigger words with relative success, they still attribute a significant amount of their importance to the wrong tokens.

Keywords: Irony Detection, Sarcasm, Explainability, Social Media, Annotation

1. Motivation & Related Work

Irony and sarcasm are often used as rhetorical devices in a face-protecting communication strategy (Brown and Levinson, 1987). When uttering criticism or refuting someone's idea, people may use positive wordings ironically as an indirect way to express their disapproval or a negative attitude. As such, ironic statements often convey the opposite of what is actually intended (Wilson and Sperber, 2012). In literature, the terms *sarcasm* and *irony* are generally used to describe the same phenomenon, although sarcasm is sometimes considered a variant of irony that is intended to offend, hurt or ridicule someone or something (Clift, 1999; Joshi et al., 2017). In this paper, we use the term *irony* for both ironic and sarcastic utterances.

In the most evident examples of irony, the author expresses a negative viewpoint using a positive wording, by stating for instance "I love it when my paper gets rejected based on one bad review". While the actual feeling of the author is not being expressed (i.e. they are not amused with the fact that the paper got rejected), the intended readers should understand the irony because they are familiar with the situation and understand the opposition that is expressed. Relying on shared implicit background knowledge makes irony often challenging to recognize, not only for humans but even more so for automated systems that do not possess such information.

To address this issue, researchers have aimed to model irony as a contrast between the sentiments of evaluations and their corresponding topics. This has been implemented both through rule-based systems (Riloff et al., 2013) and by exploring data-driven approaches (Van Hee et al., 2018c; Maladry et al., 2023b). However, others have moved away from the evaluation versus topic interpretation and instead search for text spans or "activators" that contradict each other in other ways than the sentiment they carry, such as factual oppositions and paradoxes (Karoui et al., 2017; Cignarella et al., 2018). This liberal interpretation allows them to describe more varied forms of irony. Still, both interpretations presume that annotators can confidently decide on a binary label since they possess all relevant contextual information and shared implicit knowledge. However, this is often not the case. Irony is ambiguous by design and humans may therefore doubt whether a statement is ironic or not. Forcing a binary choice on an annotator in doubt can then result in a suboptimal label that does not accurately present the annotator's opinion (Hutt et al., 2013).

LREC-COLING 2024, pages 8372–8382
20-25 May, 2024.

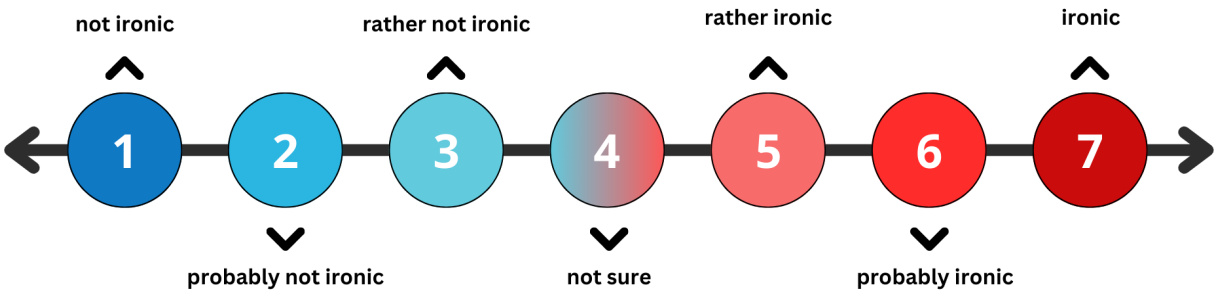

Figure 1: Representation of the irony likelihood scale.

Therefore, we propose the following novel contributions: **(1)** Enhancing the pre-existing binary irony labels with confidence information using a 7-point Likert scale for irony likelihood. By doing so, the provided labels are more nuanced and suitable to represent the subjective task of irony detection. In addition, **(2)** we tasked the annotators with indicating the trigger words that convince them that a text is ironic. This allows us to investigate to what extent human annotators use the same rationale to identify and justify the irony of a text. **(3)** After analyzing human agreement on both tasks, we investigate the performance of fine-tuned transformers on irony detection and evaluate them in light of our fine-grained annotations. **(4)** Finally, we gauge whether these automatic systems employ human-like rationales using Layer Integrated Gradients to signify the importance of each word for the system's classification.

2. Data Description

In our experimental setup we utilize the corpus built for the SemEval-2018 task 3 (Van Hee et al., 2018a). This is a collection of English tweets, which were posted between 01/12/2014 and 04/01/2015 and gathered using the search terms #irony, #sarcasm and #not. The not-ironic tweets in this corpus were posted by the same users and were collected without any specific search term. For the shared task, all tweets were manually annotated for binary irony classification in their original form (including the irony-related hashtags that were used to scrape them). This manual evaluation revealed that a significant 19% of the tweets containing these hashtags were, in fact, not ironic, or could not be defined as such without additional context (Van Hee et al., 2016). Additionally, in 52% of the ironic samples the irony hashtag was essential to recognize the irony according to the annotators. For training and evaluation, the hashtags were then removed, leaving the system without this essential contextual information. After conducting a manual evaluation, it became apparent that the irony hashtag had significantly impacted the manual labeling process. Simply omitting the hashtags for training and evaluation would lead to an unjust comparison between the manually annotated gold standard and the system predictions. We therefore decided to redo the annotations after omitting all irony-related hashtags. Considering that many tweets were previously labeled as "not ironic" even with the irony hashtags included, we retained the existing gold standard annotation for those tweets and only re-annotated tweets initially marked as ironic.

3. Annotation & Guidelines

For the re-annotation process and agreement study we enlisted three students of linguistics aged 18-25, as well as the principal investigator of this paper. English language proficiency and capability of detecting irony were evaluated for each annotator in advance and proved to be sufficient. The annotators were not specifically selected on demographic criteria, but their educational background is an asset when assessing the linguistic expression of irony. In what follows, we provide more details about the annotation procedure.

The first step in the annotation process is indicating how likely the text under investigation is ironic (i.e. the **irony likelihood**). By using a 7-point Likert scale, which was found to be the optimal number for subjective rating (Cai et al., 2016), annotators were able to indicate their confidence about the irony annotation. This novelty merges the binary classification task with a confidence indicator into the scalar annotation labels shown in Figure 1. As more labels generally imply a more complex annotation scheme, we were aware of potentially more annotator disagreement. Hence, when deemed necessary, the fine-grained labels could be merged to 5-point, 3-point or binary granularity.

In the second step of the annotation process, we asked the annotators to indicate the words or to-

kens that make the annotators believe the text is ironic (i.e. **trigger words**). Trigger words can only be indicated after an annotator has identified the tweet as ironic (i.e. an irony likelihood of at least 5 was assigned). To streamline the conceptual understanding of irony as well as the content of these trigger words, we instructed the annotators based on linguistic theory for irony and sarcasm (Wilson and Sperber, 2012; Riloff et al., 2013; Karoui et al., 2017). As discussed in the motivation, irony is often based on a contradiction that consists of either two literal components or one literal component that contradicts with the knowledge we have from the context or with common sense. Therefore, we instructed the annotators to pay specific attention to these contrasting components, which are also referred to as *markers* (Karoui et al., 2017) or *irony activators* (Cignarella et al., 2020). Of course, not all expressions of irony consist of lexicalized contrasts. More subtle ironic comments can contain absurd nonsensical statements and gross understatements or exaggerations, as well as generally inadequate reactions. Gino D'Acampo's famous reaction "If my grandmother had wheels, she would have been a bike" is a great example of such a reaction. In short, trigger words can express a contradiction, an exaggeration, a rhetorical question, a false assertion or any other type of creative expression. In addition to these conceptual guidelines, we provided liberal instructions regarding the format of trigger annotations. In essence, there were no restrictions concerning the length nor the syntactic structure of trigger words. Our only directive was for annotators to keep linguistic units, such as phrases and clauses, intact and to omit redundant specifications. In the example "I love it when my paper gets rejected based on one bad review", the subject "I" and the specification that it is "my" paper are not of significant importance. Similarly, the annotator may consider the reason of the rejection irrelevant to decide whether this tweet is ironic. Finally, punctuation could be designated as irony triggers if they conveyed semantic meaning rather than serving a purely grammatical function. Prominent examples are the expression of frustration or impatience through combined or flooded punctuation marks, such as"?!", "!!!" or "...".

4. Annotator Agreement Study

To assess the validity of our annotation scheme, we conducted an annotator agreement study on a subset of 200 tweets annotated by all four annotators. As mentioned in Section 2, we only re-annotated tweets that were labeled as ironic (when containing the irony hashtags) for the SemEval-2018 data set. Considering that the tweets were collected using an irony-related hashtag, we expect the relative number of ironic instances to be high in this data set. Furthermore, given that previous annotation found those hashtags are required to identify the irony in 52% of the tweets, the re-annotation of this set is inherently challenging.

4.1. Irony Likelihood

Table 1 displays the agreement scores for the irony likelihood annotations on a 7-point Likert scale. For completeness, we calculated the correlations with Fleiss' Kappa (Fleiss, 1971) and Krippendorff's Alpha (Krippendorff, 2011) for all annotators, as well as pairwise Cohen's Kappa (Cohen, 1960) and AC2 (Gwet, 2011) scores.[1] All agreement scores, except for Cohen's kappa, show a high agreement for initial annotation on a 7-point scale.[2] Consequently, we infer a moderate to good agreement for the irony likelihood estimation task.

Merging the fine-grained labels into fewer coarse-grained categories gradually reduces the agreement scores, as shown in Table 1. When moving from a 7-point scale to a 5-point scale, we chose to merge the outer categories "probably not ironic" (1) + "definitely not ironic" (2) and "probably ironic" (6) + "definitely ironic" (7). We reduced the 7-point scale to a 3-point scale in two ways. Option 1 (row 3) merges the outer categories (1 + 2 + 3 = not ironic; 5 + 6 + 7 = ironic) and leaves a single doubt label: "not sure" (4). Option 2 (row 4) merges labels 3, 4, and 5 (probably not + not sure + probably ironic) into a single doubt category and considers the two outer categories left and right as not ironic and ironic, respectively. The agreement results suggest that Option 2, combining the inner labels into one single doubt category, is the preferable solution for three-label granularity. Additionally, we combined the labels into two categories, following the typical approach in related research for annotating and classifying irony. This simplifies irony detection into a binary classification task. Based on the labeling scheme, this leaves the choice whether label 4 (not sure) needs to be included in the "ironic" (row 5) or the "not ironic" (row 6) category. The agreement scores indicate that it is more advisable to interpret label 4 as "not ironic" due to the significant decrease in Cohen's kappa, which offsets the 3% gain seen in the other metrics.

[1] calculated using the irrCAC package (https://github.com/kgwet/irrCAC)

[2] As shown in recent work, agreement metrics can respond differently to label imbalance (Vach and Gerke, 2023). Whereas Cohen's kappa tends to elevate with a skewed label distribution favoring the lower scales, AC2, Fleiss' Kappa and Krippendorrf's Alpha do the opposite.

Figure 2 illustrates that individual annotators exhibit different labeling patterns, with some being inclined to select confident labels more frequently, while others tend to exercise more caution, using the labels "rather (not) ironic" and "probably (not) ironic" more frequently. Gradually merging the outer labels results in a more consistent distribution among all annotators. The binary scale shows that, across all 4 annotators, an average of 29% of ironic tweets could no longer be identified as ironic. Compared to the initial SemEval study, where all of these tweets were annotated as ironic and contained irony hashtags, this indicates that the annotators overestimated the importance of the irony hashtag.

labels	F. κ	Kripp. α	C. κ	AC2
7-point	**.89**	**.89**	**.42**	**.89**
1 2 | 3 | 4 | 5 | 6 7	.85	.85	.40	.85
1 2 3 | 4 | 5 6 7	.78	.78	.32	.78
1 2 | 3 4 5 | 6 7	.85	.85	.40	.85
1 2 3 | 4 5 6 7	.77	.77	.20	.77
1 2 3 4 | 5 6 7	.74	.74	.38	.74

Table 1: Agreement scores for the irony likelihood annotation across 4 annotators. The 7-point scale and all non-binary label merging strategies use ordinal weighting.

4.2. Trigger Word Annotation

To facilitate the comparison of trigger word annotations, we first tokenize all tweets by splitting on white spaces and punctuation characters before converting them to vectors of binary values. In these binary vectors, the ones indicate the presence of trigger words, as demonstrated in Example 1.

Example 1

I	*love*	*getting*	*my*	*papers*	*rejected*	*:')*
0	*1*	*0*	*0*	*1*	*1*	*1*

Before delving into the agreement study, we provide some general statistics about the trigger word annotations in Table 2. Out of the 200 samples used for the agreement study, the annotators identified between 110 and 158 of them as ironic, with a likelihood of at least 5 on the 7-point scale. Trigger words were only indicated for these samples. Overall, these statistics indicate that the annotation of trigger words was conducted consistently, with just one annotator selecting a higher average of 10.63 trigger words per tweet.

Related work on the agreement for span annotation has observed that metrics that rely on exact span edges tend to underestimate the agreement for tasks where the span border can be subjective (Jacobs and Hoste, 2022; Lee and Sun, 2019).

	Average #Triggers	Ironic Without Triggers	Total ironic
Ann. 1	10.63 (62%)	1	153
Ann. 2	8.90 (46%)	24	110
Ann. 3	9.19 (53%)	14	158
Ann. 4	8.82 (50%)	22	147

Table 2: Average number of trigger words per annotator for each tweet (relative to tweet length between brackets) along with the number of ironic tweets without trigger words and the total number of ironic tweets. The average tweet length is 18 tokens.

Given that subjectivity is pertinent to our task as well, a strict measure of span agreement would lead to underestimation and misrepresentation of the actual agreement. Consequently, we calculate a variety of metrics with increasing degrees of stringency. All metrics were calculated pairwise and averaged over all annotator pairs for an overall agreement score.

The first evaluation criterion describes for how many texts the trigger word annotations exhibit any intersection, i.e., at least one word was indicated as trigger word by both annotators. As shown in Figure 3, this is the case for 78% of agreed-upon ironic tweets. Second, we consider the recall, which shows how many of the identified trigger words are shared with the other annotator. This shows that an average of 59% of indicated trigger words were shared in the pairwise comparisons. These two scores further confirm that there is a general consensus on the trigger words, and the discrepancy between them supports the hypothesis that the border of the larger trigger spans are more subjective.

Next, we utilize Hamming similarity to describe the entire vectors (cf. Example 1).[3] The average similarity of 64%, also in Figure 3, indicates reasonable agreement, considering the subjective nature of the task. Finally, we apply two metrics that penalize the abundance of trigger words labels: Mean-Average Precision (Beitzel et al., 2009) and Jaccard similarity coefficient. The average Mean Average Precision (MAP) of 62% and Jaccard similarity of 46% may seem modest, but these scores, along with their notable standard deviations, align with the expectations for this subjective task.

Altogether, we investigated a wide variety of agreement measures for this task and thereby present a nuanced evaluation of the trigger word

[3]Whereas Hamming distance measures the edit distance between the two vectors of binary values, we assume this distance to be relative to the entire vector length (i.e., number of tokens in a tweet) and convert this into a similarity measure by taking 1 minus the Hamming distance.

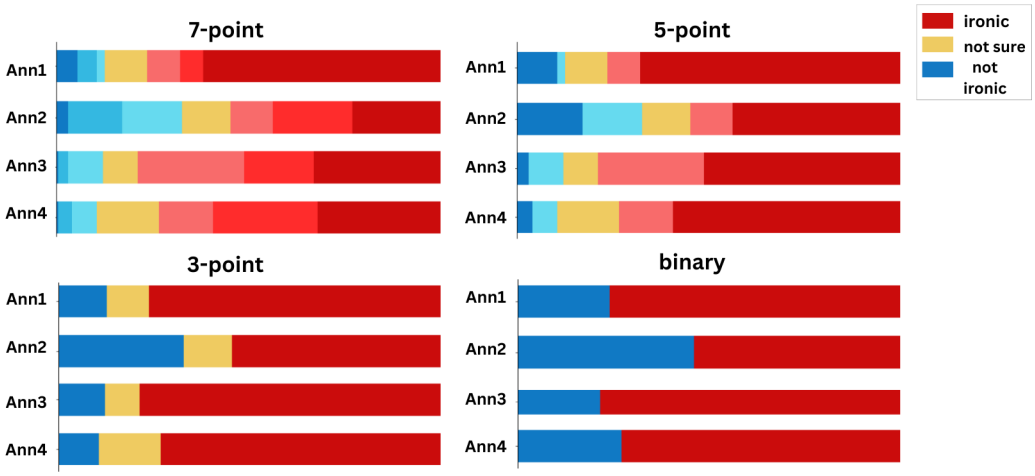

Figure 2: Label distributions for each annotator according to different merging approaches (i.e. 7-point scale to binary irony distinction).

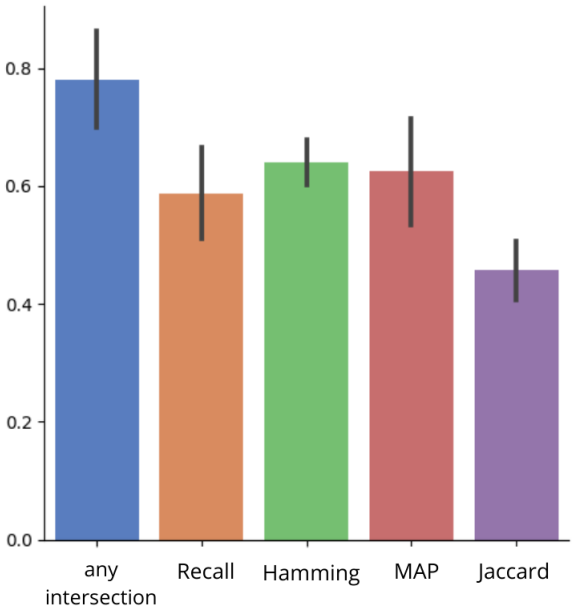

Figure 3: Average of the pairwise agreement scores and standard deviation for trigger word annotations across 4 annotators on subsets that both annotators indicated as ironic.

annotation task. In the next section, these scores will be used as an upper bound to compare to explanations generated with our automatic systems.

5. Experiments

5.1. Irony Classification

As our agreement study supports the validity of our novel annotation scheme, we proceed to explore how these annotations affect automatic systems. For our experimental corpus, we combine our fine-grained annotations on the ironic tweets with the remaining non-ironic tweets.

The label distribution for this complete annotated corpus, performed by Annotator 1, is presented in Table 3. While this comprehensive annotation proves to be insightful, all benchmark systems are trained for binary classification. For comparison, we also train a system for binary classification after merging our fine-grained labels. As discussed in Section 4, it seems preferable to consider the labels 1-4 as not ironic and 5-7 as ironic, which results in a total of 3060 (64%) genuine or not ironic tweets and 1732 (36%) ironic tweets.

	1	2	3	4	5	6	7
n	2,778	26	179	77	271	437	1,024

Table 3: Distribution of the tweets (4,792 in total) according to the 7-point scale of irony likelihood.

With these 4,792 binary labels, we fine-tuned two language and domain-specific language models: BERTweet (Nguyen et al., 2020) and Twitter-RoBERTa or T. RoBERTa (Barbieri et al., 2020). For training and evaluation, we used the entire corpus with stratified 10-fold cross-validation.[4] All models for this study are fine-tuned for 5 epochs with a batch size of 8, 200 warm-up steps, evaluating every 200 steps using a learning rate of 4e-5, optimizing with AdamW (with weight decay set to 0.04).

In addition, we compare the results to a similar model trained on the original SemEval annotations. The main difference between the SemEval gold labels and the ones in our study, is that the former were assigned based on the presence of

[4]These models are publicly available on Hugging Face through Amala3/TRoberta_Irony and Amala3/BERTweet_Irony.

irony hashtags, whereas all irony-related hashtags were discarded from our data set. Although the SemEval annotators considered irony hashtags essential to recognize the irony in 52% of the tweets (Van Hee et al., 2016), our re-annotation without the hashtags showed that only 28% of ironic tweets could not be identified as such without the irony hashtag, which means that the annotators overestimated the hashtag importance. For fair evaluation, we exclude these tweets since, as mentioned before, they were ironic in their original form, but can no longer be recognized as such.

As shown in Table 4, BERTweet slightly outperforms T. RoBERTa for both annotation setups. While the macro F1-scores are similar, our new annotation approach does outperform the system based on the original SemEval annotations by 1% macro F1 and 1.3% micro F1. However, the label-specific scores show that our annotations (of the same data) result in higher scores on the not-ironic label and lower scores on the ironic label. We hypothesize that this is connected to the minor label imbalance (64-36) we introduced with our approach compared to the balanced SemEval distribution.

		T. RoBERTa	BERTweet
Our	not ironic	.8261	**.8313**
	ironic	.7460	**.7569**
	macro	.7860	**.7941**
	micro	.7936	**.8008**
SemEval	not ironic	.7855	.7981
	ironic	.7588	.7747
	macro	.7722	.7864
	micro	.7730	.7870

Table 4: F_1-scores for fine-tuned BERT and Roberta models using 10-fold CV on the SemEval annotation setup, with a completely balanced distribution and our annotation setup, with 64% not ironic and 36% ironic distribution.

Additionally, we evaluated the best performing system (BERTweet) in the light of our fine-grained annotations. As opposed to the first evaluation, we now include the 665 tweets that were identified as not ironic during re-annotation. As discussed in section 4 and Figure 2, some annotators have a stronger tendency to use the extreme labels (1 and 7) than others. To overcome this inconsistency, we use the 5-point scale for this analysis. This means the labels are grouped as follows: as 1+2 (high confidence not ironic), 3 (low confidence not ironic), 4 (not sure), 5 (low confidence ironic) and 6+7 (high confidence ironic). In Figure 4, we display the accuracy for the labels on this five-point scale. These results show that the system performs better on the high confidence labels (1+2 and 6+7) and obtains lower scores on low confidence labels (3 and 5).

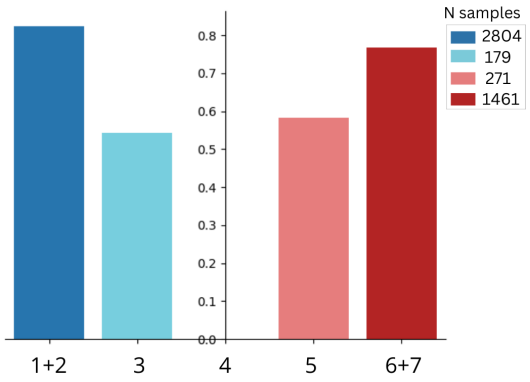

Figure 4: Accuracy on the fine-grained labels.

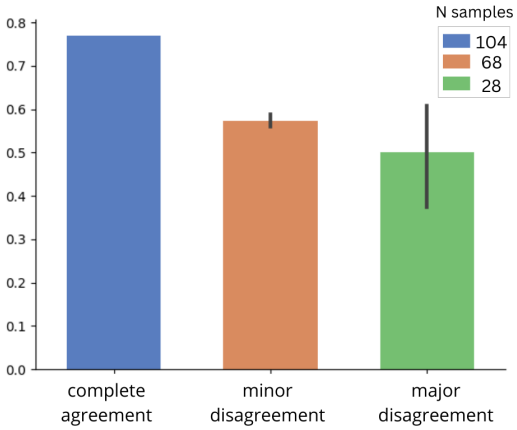

Figure 5: Accuracy for increasing levels of disagreement.

Finally, we further investigate how our system performance relates to annotator disagreement by evaluating on our agreement set as held-out test. To this end, we trained a separate model for which we excluded this set from the training data. To reflect annotator disagreement, we propose the following degrees of disagreement based on the binary merging setup: complete agreement: all annotators chose the same label, minor disagreement: a single annotator did not select the majority label, major disagreement: the annotators are divided and there is no majority label. We then proceeded to calculate the accuracy for each of the annotators and present the averaged scores in Figure 5. These results show that the system achieves lower accuracy on samples where the annotators disagree on the binary label. The standard deviation (vertical line) for the major disagreement category indicates that the system models irony in a way that is more similar to some annotators than to others.

5.2. Token Importance Comparison

One of the novelties we propose, is the annotation of irony triggers. These are the linguistic expressions that make annotators believe a text is ironic. In this section, we explore to what extent we can extract similar information from the trained models, which would allow us to determine whether fine-tuned language models employ reasoning that resembles human rationale in identifying irony. To address this research question, we utilize the held-out test set of 200 tweets and the same system we used for the held-out test scenario in the classification experiments. This allows us to compare the model's predictions to those of multiple annotators, and hence to enhance the robustness of our analysis.

In this study, we employ sub-token importance attributions utilizing Layer Integrated Gradients[5] to represent the system's trigger words. We do so, because gradient-based explanations were found to perform best for a variety of tasks (Atanasova et al., 2020) and because this approach was found to yield somewhat intuitive explanations for irony in Dutch (Maladry et al., 2023a). We implement this importance attributions metric with the transformers-interpret package[6] .

To compare these sub-token numerical importance attributions to binary human token-level attributions, several processing steps are necessary. Firstly, we aggregate multiple sub-tokens that form a single word or token, so we can compare them to human word-level annotations. Secondly, we disregard negative attributions that would indicate a word would make a tweet "not ironic". This choice aligns with the intuition that a single word cannot render a tweet non-ironic. Thirdly, we convert the raw token-level attributions into relative values between 0 and 1. In Example 2, we illustrate the effect of the aforementioned normalization steps (excluding the merging of sub-tokens into word tokens). Here, HUM. indicates the human binary annotation, whereas INIT. indicates the initial importance attributions before normalization (NORM.).

Example 2

	love	*getting*	*my*	*papers*	*rejected*	*:')*
HUM.	*1*	*0*	*0*	*1*	*1*	*1*
INIT.	*.66*	*-.06*	*.09*	*.36*	*.38*	*.49*
NORM.	*.33*	*0*	*.05*	*.18*	*.19*	*.25*

After conversion into a more uniform and intuitive format, we can now evaluate how well these token importances align with human-indicated trigger words.

Approach 1. First, we introduce a novel approach for comparing the (normalized) numerical vector to the human binary vector. For this approach, which we will call Accumulated Precise Importance, we sum the (normalized) numerical values associated with the human-identified trigger words. The goal of this approach is to evaluate how well the numerical system attributions align with human trigger words. As displayed in Table 5, this method reveals that, on average, the fine-tuned transformer models only assign 53% of the total importance scores to tokens that are important to humans. Although the scores for 3 of our golden standards are very similar, the maximum sum of 66% is significantly higher than the average. This score was achieved for the golden standard of Annotator 1, who provided the system's training labels. This discrepancy suggests that, regardless of classification performance, the system's rationale is (more than 20%) more similar to the annotator of the train labels.

	Avg.	Min.	Max.
T. RoBERTa	.525	.359	**.666**
BERTweet	**.526**	**.405**	.653

Table 5: Accumulated Precise Importance scores for the two fine-tuned systems (averaged across the four annotators and including the minimum and maximum values) for Approach 1.

Approach 2. Whereas Approach 1 describes how well the scores align with human agreement, this still leaves the question "How well does this human-system agreement compare to inter-human agreement?". To answer this question, we convert the system's numerical vector to a binary vector, which allows for a direct comparison with the human-identified trigger words. We propose two methods for this conversion. For the first method, we apply a pre-defined importance threshold. As such, all tokens with attributions composing at least x% of the total importance attribution in the sentence are considered trigger words and get the value '1', while tokens with an attribution below x% receive the value '0'. Example 3 illustrates the attribution conversion with two different thresholds applied: x = 20% and x= 10%, respectively.

Example 3

	love	*getting*	*my*	*papers*	*rejected*	*:')*
HUM.	*1*	*0*	*0*	*1*	*1*	*1*
NORM.	*.33*	*0*	*.05*	*.18*	*.19*	*.25*
x=20%	*1*	*0*	*0*	*0*	*0*	*1*
x=10%	*1*	*0*	*0*	*1*	*1*	*1*

[5]This is an approach for system interpretability based on Integrated Gradients (Sundararajan et al., 2017) that accounts for feature impact throughout different model layers.

[6]https://github.com/cdpierse/transformers-interpret

We experimented with thresholds ranging between 10% and 1%. Applying a threshold of 10% resulted in a modest average of 3 trigger words per tweet (compared to an average gold standard of 9 trigger words). Hence, we do not present the scores for thresholds above 10%.

The second method for performing the conversion involves assigning 1-values to the top-n tokens with the highest attributions. In this scenario, we set n to match the number of trigger words as identified in the corresponding gold standard, as presented in Example 4, where the first vector is the human annotation. As this approach is designed to match the exact number of trigger words of the human annotation, we assume this as the ideal scenario for Approach 2. However, unlike Approach 1, this cannot be implemented as a fully automatic approach.

Example 4

	love	*getting*	*my*	*papers*	*rejected*	*:')*
HUM.	*1*	*0*	*0*	*1*	*1*	*1*
NORM.	*.33*	*0*	*.05*	*.18*	*.19*	*.25*
TOP-N	*1*	*0*	*0*	*1*	*1*	*1*

In Figure 6, we present the agreement scores for Approach 2, where binary human vectors are compared to binary system vectors by utilizing importance thresholds on the one hand, and the number of human-indicated trigger words on the other. We employ the same metrics as used to calculate the inter-rater agreement (Section 4) and include those as an upper bound for comparison.

Our analysis of the scores, which combine the gold standards for all 4 annotators, shows that the automatic systems are very likely to recognize at least one of the human-indicated trigger words (any intersection). As expected, lowering the importance threshold improves the recall and Jaccard similarity significantly. Around the token importance thresholds of 2% the system indicates about as many tokens as trigger words as the human annotators. Although the topN approach uses additional information about the gold standard, this only results in improved Hamming similarity compared to the threshold approaches and results in lower recall scores. While all scores, with the exception of *any intersection*, are lower than human performance, the difference with this upper bound is not too large. This indicates that, given the right thresholds, automatic systems can identify trigger words with relative success, considering the subjectivity of this task. The large standard deviation in the scores for our automatic system is mainly caused by the higher results for a single Annotator. As the system was trained on annotations by Annotator 1, the system reaches significantly higher scores on their gold standard. This suggests that the system models a rationale that is more similar to the annotator it was trained for.

The collective results for these two approaches suggest that our systems can identify trigger words, but fail to estimate the importance quantitatively (see the low Accumulated Precise Attribution). On the one hand, this could be connected to the accuracy of the attribution metric, which remains hard to untangle from the system importance. On the other hand, it could be the case that the system fails to quantify the importances correctly because it is overly reliant on positive sentiment words and intensifiers, as was suggested by recent work for irony detection on Dutch (Maladry et al., 2023a,b).

6. Conclusion

In this study on the English SemEval 2018 irony data set, we proposed a fine-grained annotation scheme (3) that allows us to answer the questions “Is the text ironic?”, “How certain are you about that assessment?” and “Why do you think the text is ironic?”. Our analysis indicates strong agreement for the irony likelihood annotation and decent agreement for trigger word annotation (4). In addition, we investigated the performance of automatic systems on the same tasks. We evaluated this by checking the performance of fine-tuned transformer models for binary classification in light of our fine-grained labels (5.1) and by exploring to what extent the importance attributions of those systems align with human-identified trigger words (5.2). The evaluation of system performance revealed that our fine-grained annotations allow for slightly improved modeling of irony in automatic systems. In addition, our fine-tuned systems perform better on high-confidence and high-agreement samples compared to samples annotated with a lower confidence and provoking more inter-rater disagreement. The evaluation of trigger word detection suggests that automatic systems still assign a significant proportion of the total attribution to non-trigger word tokens. However, once the appropriate thresholds for minimum importance are identified, these systems can become more successful at identifying the important words in an ironic utterance and can reach agreement scores that are closer to human IAA scores.

For future work, our fine-grained annotations could be used to train a regression system that predicts the irony likelihood. Using these nuanced labels, the system output becomes more meaningful for a final user or when incorporated into systems for sentiment or emotion analysis. Additionally, it would be relevant to investigate what token types (emoji’s, punctuation, adjectives or adverbs) are assigned high importances and evaluate how this compares to human-identified trigger words.

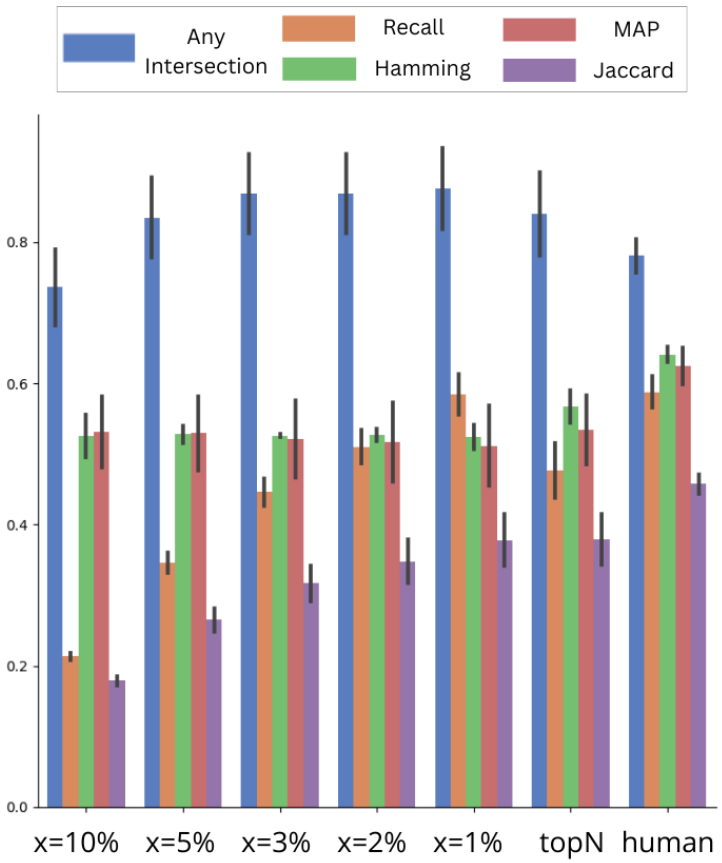

Figure 6: Agreement results for Approach 2. The figure is based on the results for BERTweet, but the results are almost identical for T. RoBERTa. Note that *human* presents the IAA scores from Figure 3.

Limitations

As irony is inherently subjective, and we only have access to limited data context, our approach is limited to making approximations regarding whether the tweets are genuinely ironic. In this paper, we do not directly address the matter of context requirements. Instead, we focus on identifying irony in a constrained information setting, devoid of conversation context or knowledge about the tweet author's background, which is valuable information for humans.

The primary limitation of this paper is the fact that we do not provide irony annotations of the full corpus for multiple annotators. Moreover, the subjectivity and complexity of the trigger word analysis didn't allow us to calculate a true statistical correlation between machine-based token importances and human-annotated trigger words for irony. Additionally, the results of our study are tied to the specific data set, which consists of a specific type of irony realized within a specific genre (i.e. microblogs) and style (i.e. with a limited number of characters and often self-indicated using hashtags). Ideally, one would prefer to train, and especially evaluate using a corpus containing ironic samples as they are found "in the wild".

Ethical Considerations

In addition to the authors of this article, student workers were employed for the annotation work. All hired workers have received a contract and a monetary compensation and for their work. As suggested by previous work (Loakman et al., 2023), it is good practice to report on the characteristics of the annotators to identify potential bias in the annotations. Therefore, we include the following description: the annotators involved are three male students in linguistics and the first author of this paper. Three out of four annotators have Dutch as their native language and the fourth annotator is a native speaker of Italian.

Acknowledgements

This work was supported by Ghent University under grant BOF.24Y.2021.0019.01 and by the Research Foundation – Flanders (FWO) under a grant for a scientific stay in Flanders (V506323N).

Bibliographical References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Steven M Beitzel, Eric C Jensen, and Ophir Frieder. 2009. Map. *Encyclopedia of Database Systems*, pages 1691–1692.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Penelope Brown and Steven C Levinson. 1987. *Politeness: Some Universals in Language Usage*. Politeness: Some Universals in Language Usage. Cambridge University Press.

MY Cai, Y Lin, and Wen-Jun Zhang. 2016. Study of the optimal number of rating bars in the likert scale. In *Proceedings of the 18th international conference on information integration and web-based applications and services*, pages 193–198.

Noam Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Alessandra Teresa Cignarella, Cristina Bosco, Viviana Patti, and Mirko Lai. 2018. Application and analysis of a multi-layered scheme for irony on the Italian Twitter Corpus TWITTIRÒ. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Alessandra Teresa Cignarella, Cristina Bosco, Viviana Patti, et al. 2017. Twittiro: a social media corpus with a multi-layered annotation for irony. In *CEUR Workshop Proceedings*, volume 2006, pages 1–6. CEUR.

Alessandra Teresa Cignarella, Manuela Sanguinetti, Cristina Bosco, Rosso Paolo, et al. 2020. Marking Irony Activators in a Universal Dependencies Treebank: The Case of an Italian Twitter Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 5098–5105. European Language Resources Association (ELRA).

Alessandra Teresa Cignarella, Manuela Sanguinetti, Cristina Bosco, Paolo Rosso, et al. 2019. Is This an Effective Way to Annotate Irony Activators? In *Proceedings of the 6th Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481, pages 1–6. CEUR-WS.

Rebecca Clift. 1999. Irony in conversation. *Language in society*, 28(4):523–553.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Kilem L Gwet. 2011. On the krippendorff's alpha coefficient. *Manuscript submitted for publication. Retrieved October*, 2(2011):2011.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hugo Hutt, Richard Everson, Murray Grant, John Love, and George Littlejohn. 2013. How clumpy is my image? Evaluating crowdsourced annotation tasks. In *2013 13th UK Workshop on Computational Intelligence (UKCI)*, pages 136–143. IEEE.

Gilles Jacobs and Véronique Hoste. 2022. Sentivent: enabling supervised information extraction of company-specific events in economic and financial news. *Language Resources and Evaluation*, 56(1):225–257.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

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.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. *Computing*, 1:25–2011.

Grace E Lee and Aixin Sun. 2019. A study on agreement in pico span annotations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1149–1152.

Tyler Loakman, Aaron Maladry, and Chenghua Lin. 2023. The iron(ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6676–6689, Singapore. Association for Computational Linguistics.

Aaron Maladry, Els Lefever, Cynthia Van Hee, and Veronique Hoste. 2023a. A fine line between irony and sincerity: Identifying bias in transformer models for irony detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 315–324.

Aaron Maladry, Els Lefever, Cynthia Van Hee, and Véronique Hoste. 2023b. The limitations of irony detection in dutch social media. *Language Resources and Evaluation*, pages 1–32.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, and David Eyers. 2021. Detecting the target of sarcasm is hard: Really?? *Information Processing & Management*, 58(4):102599.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Werner Vach and Oke Gerke. 2023. Gwet's ac1 is not a substitute for cohen's kappa – a comparison of basic properties. *MethodsX*, 10:102212.

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'16)*, pages 1794–1799.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018a. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Cynthia Van Hee, Els. Lefever, and Véronique Hoste. 2018b. We usually don't like going to the dentist : using common sense to detect irony on twitter. *Computational Linguistics*, 44(4):793–832.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018c. We usually don't like going to the dentist: Using common sense to detect irony on twitter. *Computational Linguistics*, 44(4):793–832.

Byron C Wallace, Laura Kertz, Eugene Charniak, et al. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516.

Wilson and Sperber. 2012. Explaining irony. *Meaning and relevance*, pages 123–145.

Language Resource References