Context vs. Human Disagreement in Sarcasm Detection

Hyewon Jang[1], **Moritz Jakob**[1], **Diego Frassinelli**[1,2]
[1]Department of Linguistics, University of Konstanz, Germany
[2]Center for Information and Language Processing, LMU Munich, Germany
{hye-won.jang, moritz.jakob, diego.frassinelli}@uni-konstanz.de

Abstract

Prior work has highlighted the importance of context in the identification of sarcasm by humans and language models. This work examines *how much* context is required for a better identification of sarcasm by both parties. We collect textual responses to dialogical prompts and sarcasm judgment to the responses placed after long contexts, short contexts, and no contexts. We find that both for humans and language models, the presence of context is generally important in identifying sarcasm in the response. But increasing the amount of context provides no added benefit to humans (long = short > none). This is the same for language models, but only on easily agreed-upon sentences; for sentences with disagreement among human evaluators, different models show different behavior. Also, we show how, despite the low agreement in human evaluation, the sarcasm detection patterns by the manipulation of context amount stay consistent.

1 Introduction and related work

This work examines the role of the presence and amount of contextual information in detecting sarcasm. Previous work in cognitive science has shown the importance of context in sarcasm comprehension (Woodland and Voyer, 2011) and production (Jang et al., 2023) for humans. In computational linguistics, similar observations were made: supplying context to the target utterance boosts sarcasm detection performance of language models, though with more conflicting results: some studies report that supplying context leads to a performance boost in sarcasm detection by neural models (Jaiswal, 2020; Ghosh et al., 2018), whereas other studies report no such benefit (Castro et al., 2019) or marginal benefit (Jang and Frassinelli, 2024) in using context for the same task. However, there has not been much effort in exploring the benefit of varying amounts of contextual information, or in addressing what counts as context. The term 'context' varies a lot work by work; it can mean any number of preceding strings such as previous posts on social media (Jaiswal, 2020; Joshi et al., 2016) or previous utterances in a dialogue (Castro et al., 2019), or any additional information that can help detect sarcasm, such as eye-tracking data (Mishra et al., 2016) or images (Schifanella et al., 2016).

In this work, we define context as the preceding textual utterances that can trigger sarcasm in people (Section 2), and then examine what is a good amount of contextual information that facilitates sarcasm identification for humans (Section 3) and language models (Section 4). We further show how context interacts with the level of disagreement among human evaluators (Section 4.3).

2 Data creation

We created a new dataset based on the Multimodal Sarcasm Detection Dataset (MUStARD; Castro et al., 2019). The MUStARD dataset contains written transcriptions of "contexts" (preceding utterances) and the following "response"[1] from multiple TV series, and binary labels of sarcasm for the responses (*sarcastic* or *not sarcastic*). We selected 24 contexts that are generalizable enough, all of which were from the TV series 'Friends' and situations happening between two conversation partners. The names of all conversation partners were modified to detach the stimuli from the TV show as much as possible. For all the selected contexts, we collected new responses in an online data collection.

Here, we manipulated the amount of context. Additional to the original contexts available in a short utterance form, we described each context in a narrative form by manually referring to the scenes and episodes of the TV show to restore the relevant information that would allow the following utterance to be correctly judged as sarcastic or not. This information in the original dataset often came

[1]The term used in the MUStARD dataset is 'utterance'.

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 1–7

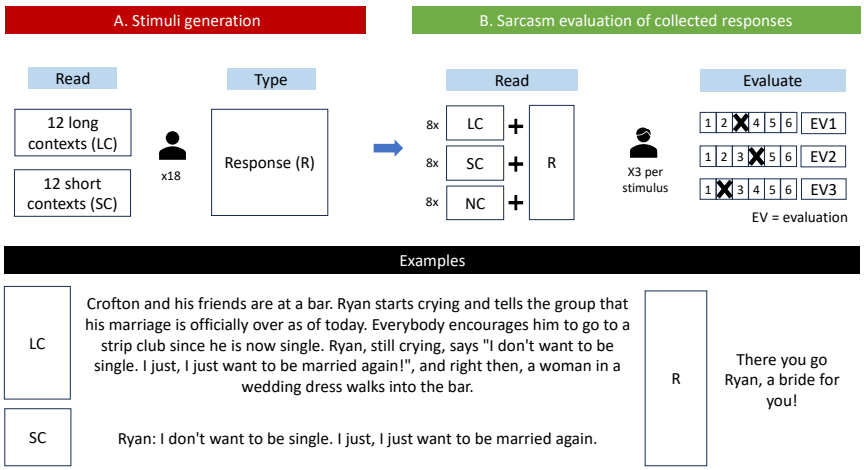

Figure 1: Data collection (A), data evaluation (B), and example stimuli for long (LC) and short (SC) contexts, and an example response (R) collected from participants.

from multimodal, episode-level, or series-level information not reflected in the transcripts.

Therefore, each context was represented twice both as *short context* (SC) in its original utterance form and as *long context* (LC) in a descriptive/narrative form. The average number of words was 26 for SC and 66 for LC. For each LC and SC, we collected new responses to make the stimuli comparable, given that the original dataset had responses only to short contexts. This also allowed us to collect spontaneous responses from multiple lay people as opposed to responses generated by professional screenwriters.

We recruited 32 native English-speaking participants based in the UK, USA, Canada, Australia, New Zealand or Ireland[2]. They read 24 contexts and freely responded to each (they were not instructed to be sarcastic). Half of the contexts (N = 12) were presented as SC and the other half as LC (See A in Figure 1). At the end of the collection, participants reported their familiarity to the TV show Friends and how many of the situations they recognized as being from the show.

To control for the expectation of sarcasm arising from the familiarity to the TV show, we discarded data from the participants who were *quite familiar*, *very familiar*, or *extremely familiar* to the show or who recognized at least 3 scenes from the show. After removing data from 14 such participants, data by 18 respondents remained.[3]

[2]We used FindingFive (https://www.findingfive.com) for experiment building and Prolific (https://www.prolific.co) for participant recruitment.

[3]The new data consisting of responses and evaluation ratings are available at https://github.com/copsyn.

3 Influence of context for sarcasm judgment by humans

Here we identify what amount of context affects human judgment of sarcasm on the following response.

3.1 Experiment

In an online experiment, new participants evaluated the level of sarcasm of the responses in isolation (NC) or placed after long context (LC) or short context (SC) as shown in Table 1.

Table 1: Number of items for different combinations of context (**C**) and response (**R**).

	Condition	N
i	SC (24) + R (18)	432
ii	LC (24) + R (18)	432
iii	NC (R-only)	432
Total		1,296

In conditions **i** and **ii**, each context is paired with the generated responses and condition **iii** consists of the responses only (See Section 2).

Each stimulus was evaluated by 3 participants recruited with the same criteria as before. Each participant was presented with 24 stimuli, distributed evenly across the 3 conditions (See B in Figure 1). Participants rated the sarcasm level of the responses on a six-point Likert scale (*not at all*, *mostly not*, *not so much*, *somewhat*, *mostly*, and *completely*). Participants who failed attention check questions

or were familiar with the TV show were replaced with new ones.

Context length and disagreement Table 2 shows the proportions of sarcasm (binary-coded from the six-point scale; *completely*, *mostly*, *somewhat* into *sarcastic*) in each contextual condition by three evaluators and by their average per stimulus. The probability of judging a response as sarcastic increases when contextual information is present. Around 38% of instances that were judged as 'not sarcastic' in the NC condition were judged as 'sarcastic' when more context became available (LC or SC condition). However, adding context also increases disagreement among evaluators (lower Kappa).

Table 2: Proportions of sarcastic responses (binary-coded) by context amount according to three distinct evaluations per stimulus (EVs) and inter-rater agreement (Fleiss' Kappa) by context amount.

	AVG	EV1	EV2	EV3	Kappa
LC	0.46	0.36	0.44	0.49	0.10
SC	0.42	0.36	0.43	0.41	0.13
NC	0.23	0.25	0.25	0.28	0.18

3.2 Analysis and results

We tested whether the presence and amount of contextual information are important factors for humans to identify sarcasm in the following response. To easily compare the behavior of humans and LMs, we binarized the sarcasm ratings. The overall inter-rater agreement across all stimuli measured by Fleiss' Kappa was 0.17 (See Appendix B for Spearman correlations).[4]

We fit a generalized linear mixed-effects model for each evaluation (See Appendix C for details)[5]. Random intercepts for participants and items were included in the statistical model. We used R (R Core Team, 2021) and the *lme4*-package (Bates et al., 2015) for the main models and the *emmeans*-package for post-hoc pairwise comparisons (Lenth, 2023).

For all evaluations, the presence of context, either long or short, triggered significantly higher probability of perceiving sarcasm in the following response. Long contexts caused more frequent sarcasm judgment compared to short contexts only in EV3 ($p < 0.005$), but not in EV1 ($p = 0.98$), EV2 ($p = 0.97$), or AVG ($p = 0.27$). The results indicate that the presence of context is important for human evaluators to identify sarcasm, but a greater amount of context does not necessarily lead to any added benefit.

[4]For comparison, the Kappa score reported in the original MUStARD paper is 0.23 (Castro et al., 2019).

[5]In this work, unless otherwise specified, statistically significant scores correspond to a p-value smaller than 0.001.

4 Influence of context on sarcasm detection by large language models

Here we test if manipulating the amount of context directly affects the performance of three language models in the detection of sarcasm on the following response. As gold standard we use the human-evaluated scores described in Section 3.

4.1 Data and model

We performed sarcasm detection using three pretrained LMs: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019). We fine-tuned these models on the contexts and responses from the MUStARD dataset excluding the 24 contexts we used in our experiments. We then used our data as test data to classify the responses in the three conditions (LC, SC, NC) as either *sarcastic* or *not sarcastic*. Given the high subjectivity in identifying sarcasm indicated by the low inter-rater agreement (Kappa 0.17), we predicted the binary-coded human ratings by the three evaluations (EVs) independently and combined. We conducted an error analysis comparing the results from the three EVs. We used four different seeds and five folds for validation. All the reported results in this paper are an average of all the models (4 seeds $\times$ 5 folds) trained for 10 epochs, which yielded the best prediction results. See Appendix A for the full model parameters.

Table 3: Macro F-scores of sarcasm detection on the new dataset described in Section 2 by three LMs trained on MUStARD for 10 epochs. Labels provided by each evaluation (EV) or combined (C) across three EVs.

		EV1	EV2	EV3	C
BERT	LC	0.49	0.52	0.57	0.55
	SC	0.53	0.51	0.53	0.54
	NC	0.47	0.39	0.41	0.36
RoBERTa	LC	0.46	0.54	0.52	0.53
	SC	0.54	0.50	0.52	0.50
	NC	0.36	0.34	0.38	0.29
DistilBERT	LC	0.53	0.52	0.51	0.53
	SC	0.53	0.51	0.52	0.53
	NC	0.44	0.38	0.40	0.32

4.2 Results

Overall, the three LMs achieve comparable classification results. Supplying context, either short or long, always improves the performance of all LMs. The performance results in Table 3 suggest that there are no strong differences between supplying long context and short context. A noteworthy aspect of these results is that despite low agreement among three evaluations, the prediction results by context amount show similar patterns for all EVs (LC and SC lead to a higher number of correct predictions than NC).

4.3 Error analysis

Disagreement among human evaluators To identify the reasons behind the similar patterns in model performance despite low agreement, we divided the data into *agreed-upon* (all evaluators agreed on a label) and *disagreed-upon* (evaluators disagreed on the label: 2 vs. 1) instances of sarcasm based on the binarized labels. From the *disagreed-upon* category, we extracted the number of instances for which LMs chose the majority label (better choice) or the minority label (worse choice), neither of which is completely correct or incorrect. Table 4 shows that LMs choose the labels given by each evaluation at a similar rate. This pattern suggests that LMs misclassify some sentences when tested with labels from one evaluation, but misclassify other sentences when tested with labels from another evaluation, thus holding the general classification patterns stable.

Table 4: Proportions (Prop.) of predictions by BERT. Correct & incorrect predictions apply to *agreed-upon* (A) instances. Majority (better choice) & minority (worse choice) predictions apply to *disagreed-upon* (D) instances. The other models show the same pattern (See Appendix D).

Type	Prediction	Evaluations that predictions match			Prop.
A	Correct	All			0.58
	Incorrect	None			0.42
		Match_EV1	**Match_EV2**	**Match_EV3**	
D	Majority	0	1	1	0.18
		1	0	1	0.17
		1	1	0	0.17
	Minority	0	0	1	0.15
		0	1	0	0.16
		1	0	0	0.17

The interaction between context amount and degree of disagreement To analyze the interaction between the amount of context (LC, SC, NC) and disagreement levels (agreed vs. disagreed), we categorized the predicted labels according to these factors. Table 5 shows that for *agreed-upon* instances, providing context helps LMs predict (more) correct labels than when no contexts are available (LC/SC >NC for correct & majority). For *disagreed-upon* instances, more variability is shown: For BERT, only long context significantly improves the detection of sarcasm (LC >SC = NC), whereas for RoBERTa and DistilBERT, no amount of context is beneficial (LC = SC = NC).

Table 5: Proportions of classification choice of BERT (average across all seeds and folds) by context length × disagreement level.

		Agreed-upon			**Disagreed-upon**		
		Correct	Incorrect	Std.	Majority	Minority	Std.
BERT	LC	0.60	0.40	0.07	0.54	0.46	0.04
	SC	0.60	0.40	0.07	0.51	0.49	0.05
	NC	0.55	0.45	0.16	0.50	0.50	0.06
RoBERTa	LC	0.61	0.39	0.09	0.50	0.50	0.04
	SC	0.57	0.43	0.08	0.48	0.52	0.04
	NC	0.52	0.48	0.14	0.49	0.51	0.06
DistilBERT	LC	0.57	0.43	0.08	0.52	0.48	0.05
	SC	0.59	0.41	0.10	0.51	0.49	0.05
	NC	0.54	0.46	0.19	0.50	0.50	0.08

In summary, the presence of context is important for LMs to significantly improve their performance of sarcasm detection for sentences with a high agreement, but adding more context does not present clear benefit compared to a lower amount of context. For sentences with disagreement, the contribution of contextual information heavily depends on each model. Only BERT uses the extra contextual information provided by a longer context to detect sarcasm significantly better.

5 Conclusion

This work systematically tested the amount of contextual information required for humans and language models to evaluate the following utterance in terms of sarcasm. We showed that in general, the presence of context leads to better detection of sarcasm both by humans and by three LMs. But, providing a higher amount of information in the context did not present clear additional benefit for humans, which was also true for LMs for sentences for which human evaluators agreed on a label. When humans disagreed, the presence of context stopped playing any role in facilitating the detection of sarcasm in RoBERTa and DistilBERT, whereas the performance of BERT improved when a longer context was provided. We lastly showed

that low inter-rater agreement did not affect the overall classification patterns, due to a high variability in the sentences that the models misclassify each time they are tested against labels from different human evaluators. This is a relevant finding for many NLP tasks prone to disagreement and susceptible to subjectivity, which must continue to be addressed in future research.

Limitations

This work investigated the influence of the amount of information embedded in the context. However, we did not systematically calculate the amount of information available in the different contextual conditions (SC vs. LC). Future work should address how to draw a line between sufficient and redundant contextual information by investigating a gradient change in the amount of context.

The data collected in this work is small because we had to go through rigorous filtering of an existing dataset to obtain sufficiently generalizable contexts for further experiments. Future work should test the same effect with a bigger sample size.

In the data collection (Section 2), we only recruited male participants because some of the selected situations were much more suitable for male speakers than female speakers and the already small number of generalizable contexts could not be further reduced. A follow-up study should include gender as a variable for a more comprehensive evaluation of the use of sarcasm by humans.

Ethics Statement

We see little ethical issue related to this work. All our experiments involving human participants were conducted anonymously, on a voluntary basis, and with a fair compensation suggested by the recruitment platform Prolific (9 GBP per hour) and are in line with the ethical regulations of the University of Konstanz (IRB number 05/2021). All our modeling experiments were conducted with open-source libraries, which received due citations. However, we acknowledge that some of the stimuli extracted from the original MUStARD dataset contain sensitive language that could potentially be insulting for the reader.

Acknowledgements

We thank Matteo Guida and Hsun-Hui Lin for their work in initial data selection and preparation.

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

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.

Debanjan Ghosh, Alexander R. Fabbri, and Smaranda Muresan. 2018. Sarcasm Analysis Using Conversation Context. *Computational Linguistics*, 44(4):755–792.

Nikhil Jaiswal. 2020. Neural sarcasm detection using conversation context. In *Proceedings of the second workshop on figurative language processing*, pages 77–82.

Hyewon Jang, Bettina Braun, and Diego Frassinelli. 2023. Intended and perceived sarcasm between close friends: What triggers sarcasm and what gets conveyed? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! *arXiv preprint arXiv:2404.06357*.

Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, and Mark J. Carman. 2016. Harnessing sequence labeling for sarcasm detection in dialogue from TV series 'Friends'. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 146–155, Berlin, Germany. Association for Computational Linguistics.

Russell V. Lenth. 2023. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.6.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016. Harnessing

cognitive features for sarcasm detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Berlin, Germany. Association for Computational Linguistics.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.

Jennifer Woodland and Daniel Voyer. 2011. Context and intonation in the perception of sarcasm. *Metaphor and Symbol*, 26(3):227–239.

A Fine-tuning implementation details

We used `bert-base-uncased`, `roberta-base`, and `distilbert-base-uncased`. Each language model was fine-tuned for 2, 5, and 10 epochs with a batch size of 64, a learning rate of 5e-5, and a weight decay of 1e-2. The fine-tuning was implemented using the Trainer class from the Hugging Face library, and conducted on an NVIDIA A100 GPU with a total memory of 40GB.

B Inter-rater agreement

Table 6 reports the Spearman's correlation coefficients (r) calculated between the original ratings (1-6 Likert scale) that each evaluation group (EV) assigned to responses alone (NC) and responses following long contexts (LC) or short contexts (SC). The trends observed here are consistent with the results on the binarized sarcasm scores reported in Table 2 in the main text.

Table 6: Inter-rater agreement of the original ratings (1-6) measured by Spearman's correlations between each pair of evaluation (EV), $p < 0.005$.

	EV1-EV2	EV1-EV3	EV2-EV3
LC	0.26	0.17	0.15
SC	0.26	0.17	0.19
NC	0.18	0.24	0.20

C Details of statistical tests

The formula used for the GLMER models is as follows:

sarcasm_binary_labels ∼
context_amount
+ (1 | item) + (1 | participant)

The model indicates if there are differences in the sarcasm label (yes/no) distribution given contextual manipulation. The random intercepts account for the variability between participants and items that cannot be explained by the fixed effects alone.

The *emmeans* library conducts a pairwise comparison of the three context conditions (LC vs. SC, LC vs. NC, and SC vs. NC) by performing automatic alpha correction.

D Error analysis for the other models

Proportions of predictions by RoBERTa (see Table 7) and DistilBERT (see Table 8).

Table 7: Proportions (Prop.) of predictions by RoBERTa. Correct & incorrect predictions apply to *agreed-upon* (A) instances. Majority (better choice) & minority (worse choice) predictions apply to *disagreed-upon* (D) instances.

Type	Prediction	Annotator groups that predictions match			Prop.
A	Correct	All			0.56
	Incorrect	None			0.44
		Match_EV1	**Match_EV2**	**Match_EV3**	
D	Majority	0	1	1	0.16
		1	0	1	0.16
		1	1	0	0.17
	Minority	0	0	1	0.18
		0	1	0	0.16
		1	0	0	0.17

Table 8: Proportions (Prop.) of predictions by DistilBERT. Correct & incorrect predictions apply to *agreed-upon* (A) instances. Majority (better choice) & minority (worse choice) predictions apply to *disagreed-upon* (D) instances.

Type	Prediction	Annotator groups that predictions match			Prop.
A	Correct	All			0.56
	Incorrect	None			0.44
		Match_EV1	**Match_EV2**	**Match_EV3**	
D	Majority	0	1	1	0.17
		1	0	1	0.17
		1	1	0	0.18
	Minority	0	0	1	0.17
		0	1	0	0.16
		1	0	0	0.16