Generalizable Sarcasm Detection Is Just Around The Corner, Of Course!

Hyewon Jang[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, diego.frassinelli}@uni-konstanz.de

Abstract

We tested the robustness of sarcasm detection models by examining their behavior when fine-tuned on four sarcasm datasets containing varying characteristics of sarcasm: label source (authors vs. third-party), domain (social media/online vs. offline conversations/dialogues), style (aggressive vs. humorous mocking). We tested their prediction performance on the same dataset (intra-dataset) and across different datasets (cross-dataset). For intra-dataset predictions, models consistently performed better when fine-tuned with third-party labels rather than with author labels. For cross-dataset predictions, most models failed to generalize well to the other datasets, implying that one type of dataset cannot represent all sorts of sarcasm with different styles and domains. Compared to the existing datasets, models fine-tuned on the new dataset we release in this work showed the highest generalizability to other datasets. With a manual inspection of the datasets and post-hoc analysis, we attributed the difficulty in generalization to the fact that sarcasm actually comes in different domains and styles. We argue that future sarcasm research should take the broad scope of sarcasm into account.

1 Introduction

Sarcasm can be used to hurt, criticize, or deride (Colston, 1997; Frenda et al., 2022; Keenan and Quigley, 1999; Kreuz and Glucksberg, 1989) but also to be mocking, humorous, or to bond (Dews et al., 1995; Gibbs, 2000; Pexman and Olineck, 2002). It comes in many different shapes such as understatement, hyperbole, rhetorical questions (Leggitt and Gibbs, 2000), deliberate falsehood (Glucksberg, 1995; Riloff et al., 2013), or self-deprecation (Abulaish and Kamal, 2018). Given such complexity and diversity of sarcasm, one should ask whether current sarcasm detection models are robust enough to be easily applied to any kind of sarcasm. This question is particularly valuable in computational linguistics, as most prior

work addressing sarcasm adopted a narrow working definition of sarcasm as 'saying the opposite of the true message (often with the intent to be hurtful)' (Cai et al., 2019; Frenda et al., 2022; Ghosh and Veale, 2017; Joshi et al., 2015; Pan et al., 2020). This definition is based on the traditional theory of sarcasm by Brown and Levinson (1978), but is not at all comprehensive (Oprea et al., 2021; Sperber and Wilson, 1981).

Prior work in sarcasm detection has used one dataset at a time that only includes a specific style and domain of sarcasm. This can lead to frail sarcasm detection models because different datasets come from different sources such as social media (Abu Farha et al., 2022; Khodak et al., 2018), online forums (Oraby et al., 2016), TV series (Castro et al., 2019), product reviews (Filatova, 2012), or conversations (Chakrabarty et al., 2022). The vast range of domains, styles, and topics that sarcasm can be associated with invites the question of whether sarcasm detection models fine-tuned on a specific dataset are generalizable to other datasets.

Also, existing datasets of sarcasm contain sarcasm labels from different sources: tags (e.g., #sarcasm or /s) (Khodak et al., 2018; Joshi et al., 2015), labels by annotators (Castro et al., 2019; Oraby et al., 2016; Riloff et al., 2013), or labels by the authors of the posts (Oprea and Magdy, 2020). Collecting data with labels based on tags is easy and scalable, but the data tend to be noisy and it poses a high risk of including false positives in the dataset (Khodak et al., 2018; Sykora et al., 2020). Most manually annotated data come with labels provided by crowd workers or experts (though it raises the question of "who is an expert of sarcasm?"). These labels are more reliable than tags, but they still pose a risk of wrongly reflecting the intention of the original utterance maker (Oprea and Magdy, 2020) as sarcasm can be missed (referred to by some as *sarchasm*) even in multimodal settings (Fox Tree et al., 2020). In text-only settings, which is what

4238

a lot of sarcasm detection models are based on, the chance of misinterpreted sarcasm would naturally be much higher as the available cues are more limited. Recent work has reported that previous state-of-the-art models fine-tuned on datasets with third-party labels showed vastly different (poorer) performance when predicting sarcasm on data with author labels, going from an F-score of 0.75 to 0.33 (Abu Farha et al., 2022; Oprea and Magdy, 2020).

Thus, in this work, we investigate the generalizability of sarcasm detection models fine-tuned on different datasets that contain sarcasm from different domains (social media/online vs. offline conversations/dialogues), styles (aggressive vs. humorous), and labels from different perspectives (author labels vs. third-party labels). Specifically, we make the following contributions:

C1. Through cross-dataset comparisons, we test the generalizability of sarcasm detection models, which have been fine-tuned on various datasets, to other datasets. We identify the datasets that lead to the highest prediction accuracy scores across datasets.

C2. We test the effect of label source (author vs. third-party) on model performance by using a new dataset that we release.

C3. We analyze different styles of sarcasm found in different datasets and argue that future work should include a more diverse shapes of sarcasm.

2 Related work

If there is any consensus around the topic of sarcasm, it is arguably the fact that sarcasm is hard to define (Fox Tree et al., 2020). But much is known about the communicative functions of sarcasm as they have been extensively studied in theoretical and experimental work. Sarcasm can be used to mock (Gibbs, 2000; Pexman and Olineck, 2002), to criticize harsher (Colston, 1997), to be humorous (Gibbs, 2000; Glucksberg, 1995; Matthews et al., 2006), or to save one's own face (Jorgensen, 1996). Sarcasm is not only multifaceted in its functions but also manifests itself in diverse forms, including hyperbole, often accompanied by interjections and intensifiers (Joshi et al., 2017), rhetorical questions (Oraby et al., 2016), violation of maxim of relevance (Sperber and Wilson, 1986) or maxim of quality (Grice, 1975).

Despite the complications surrounding this topic, a lot of work in computational linguistics came up with operational definitions of sarcasm for automatic sarcasm detection (i.e., contextual incongruity (Joshi et al., 2015), and discrepancy between positive sentiment and negative situation (Riloff et al., 2013)). Sarcasm detection has since had fluctuating success rates from an F-score of 0.51 to 0.97 (Bǎroiu and Trǎuşan-Matu, 2022).

One reason behind varying results may come from the nature of the data. Twitter is the source that most previous computational work on sarcasm has relied on as training data (Barbieri et al., 2014; Joshi et al., 2015; Ptáček et al., 2014; Van Hee et al., 2018). Other sources of sarcasm data include online forums (Oraby et al., 2016; Walker et al., 2012), Reddit (Khodak et al., 2018), or TV series (Castro et al., 2019). Data sources matter for building a more robust sarcasm detector; Joshi et al. (2015) showed that the same models performed quite differently across different datasets, with variations in F-scores of around 0.20.

More recently, the importance of addressing the sources of sarcasm labels has also been raised in multiple studies (Abu Farha et al., 2022; Oprea and Magdy, 2019, 2020). They collectively showed that sarcasm detection models perform significantly worse when making predictions on data with labels provided by the authors of sarcastic utterances. This is a factor worth examining as most datasets were created by having human annotators rate the level of sarcasm from the observer's point-of-view (Castro et al., 2019; Riloff et al., 2013; Van Hee et al., 2018).

In our work we investigate the robustness of sarcasm detection models by taking into account the label source, data source, types, and styles of sarcasm.

3 Datasets

We take four existing datasets of sarcasm and use them in our experiment. We select these datasets because they each have different elements that facilitate cross-data comparison. In Section 3.1, we provide descriptions for each dataset, followed by the strengths and limitations of each that motivate the creation of a new dataset, which we describe in Section 3.2. We provide examples from each dataset in Table 1.

3.1 Existing resources

MUStARD (The Multimodal Sarcasm Dataset) is a dataset of sarcasm that Castro et al. (2019) curated out of several TV shows such as *Friends* and

Table 1: Two examples from each of the four datasets to illustrate the different styles of sarcasm.

Dataset	Ex. No.	Examples from each dataset
CSC	1	**Context**: Steve gives you a watering can on your birthday while smiling at you with a strange expression. But you don't even have a single plant. **Response**: Maybe I will use it as an outside shower.
	2	**Context**: You know that Steve does almost everything at the last minute. Today Steve tells you that he has to write a joint proposal with colleagues from a different team by the end of this week, and complains that things are moving so slowly because everybody in the other team is a procrastinator. He says, "gosh, it's so frustrating that I'm the only one trying to get the stuff moving. Everyone in this team is too laid back!" **Response**: Oh yeah, and you're not laid-back at all. You are the emperor of procrastination, Steve! If you're the most highly motivated person on that team, they really must be the most unmotivated team in human history.
MUStARD	1	**Context**: 'How do I look?', 'Could you be more specific?', "Can you tell I'm perspiring a little?" **Response**: No. The dark crescent-shaped patterns under your arms conceal it nicely.
	2	**Context**: "So. I just thought the two of us should hang out for a bit. I mean, you know, we've never really talked." **Response**: I guess you'd know that, being one of the two of us, though, right?
SC V2	1	Aaahhh, so just not accomplishing our goals and going home is not defeat, there has to be paperwork for it to be a defeat. Gotcha.
	2	Ever hear of artificial ensemination? Why is that heteros only think there is one way to produce children? I find hetero sex disturbing, and an unnatural lifestyle choice.
iSarcasmEval	1	Imagine going to university for 4 years when you could just follow Elon Musk on Twitter for free.
	2	The control button just fell off my laptop. How symbolic for my life.

The Big Bang Theory. The dataset comes with transcripts, audio, and video files of 690 scenes. Each scene consists of contextual utterances followed by a target utterance, half of which are sarcastic (N = 345). Each scene was annotated by 2 annotators, who gave binary sarcasm labels to each video. The inter-rater agreement was a Kappa score of 0.23 for the majority of the videos and 0.59 for the remaining ones; a third annotator resolved the possible disagreement.

Strengths: MUStARD is a multimodal dataset, and the annotation was done based on both contexts and responses. Despite its multimodal nature, there is work that used the dataset in text-only settings as well (Das et al., 2023; Zhang et al., 2021).

Weaknesses: The dataset is relatively small, the annotator agreement is low, and the sentences are scripted (as opposed to naturally-occurring).

Sarcasm Corpus V2 (Oraby et al., 2016) is a dataset extracted from the Internet Argument Corpus, containing posts from 3 different online debate sites (Abbott et al., 2016; Walker et al., 2012). Sarcasm Corpus V2 contains a total of 9,386 written sentences, half of which are sarcastic. Each sentence was annotated by 9 annotators, who gave binary labels after being shown context posts with a following response. The inter-rater agreement was 0.80 (in percent agreement) among 9 annotators and 0.89 among the 3 best annotators.[1]

Strengths: Sarcasm Corpus V2 is large, and several annotators contributed to the annotation based on both contexts and responses.

Weaknesses: The dataset itself does not provide contexts, and the source domain (online debate sites) can bias the data into a specific type of language use that is not commonly found in regular conversations.

iSarcasmEval (Abu Farha et al., 2022) is a dataset containing Twitter posts and sarcasm labels by the authors of the posts. iSarcasmEval contains 3,801 sentences, 1,067 of which are sarcastic.[2] Native English speakers on Prolific who were users of Twitter provided their own previous sarcastic and non-sarcastic posts with explanations of why they thought their posts were sarcastic. As the dataset has single author labels instead of third-party labels, no inter-rater agreement is applicable.

Strengths: iSarcasmEval provides author labels, which distinguishes itself from the previous datasets.

Weaknesses: It does not contain any contextual information.

3.2 New dataset: Conversational Sarcasm Corpus (CSC)

We aimed to construct a new dataset that can overcome some of the main limitations of the existing

---

[1]The agreement is for one of the three subsets. The authors do not report the agreement score for the entire dataset.

[2]Abu Farha et al. (2022) report slightly different numbers. The numbers reported in this paper are according to their data repository at https://github.com/iabufarha/iSarcasmEval.
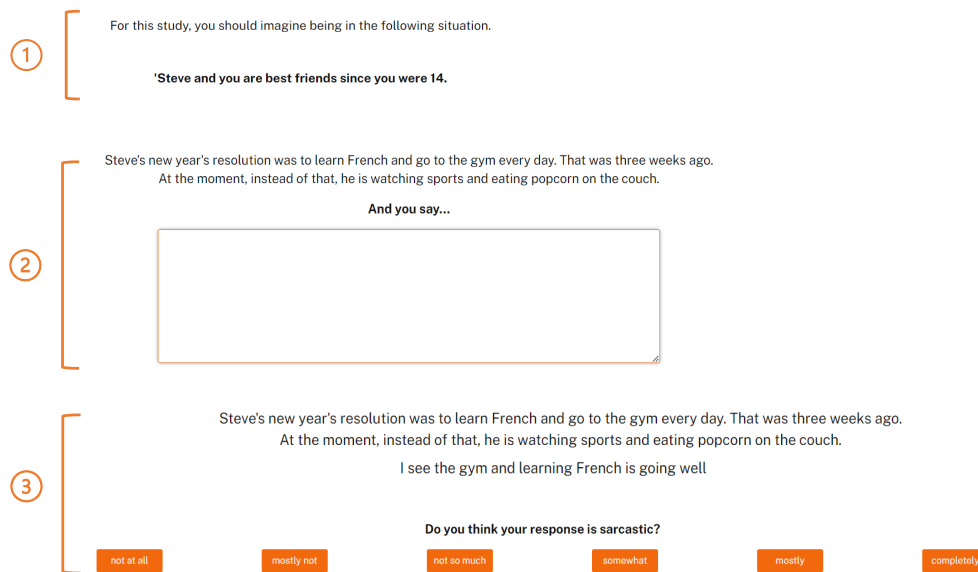
Figure 1: Data collection process illustration. Step 1 was shown to participants at the beginning of the experiment. All situations were shown as in Step 2 and afterwards, participants evaluated their own responses as shown in Step 3.

datasets. To that end, we created a new sarcasm dataset that is based on offline conversational contexts (as opposed to Sarcasm Corpus V2 and iSarcasmEval), that is larger (than MUStARD and iSarcasmEval), and that contains naturally-occurring responses by a number of people (as opposed to MUStARD). The dataset contains author labels as well as third-party labels (as opposed to all the others). We constructed the dataset in two phases (part 1 and part 2) in order to first confirm our hypothesis of certain contextual prompts being more likely to trigger sarcastic responses. We then expanded our collection using the same approach but with different contextual prompts.

In part 1[3], we created 32 situational descriptions, sometimes followed by an utterance by an interlocutor (The example in Figure 1 ends with a description, but some stimuli were followed by an additional utterance by Steve). Half of the situational descriptions were created in a way in which the interlocutor was behaving in silly or annoying manners. We assumed that such situations would trigger more sarcastic responses than others based on our qualitative analysis of the MUStARD dataset. We recruited 60 (30 female, mean age = 33) native English speakers online[4]. They were asked to imag-

ine being a close friend of the interlocutor in the 32 situations (Step 1 in Figure 1). They freely responded to all situations by typing (Step 2 in Figure 1), and after this process was finished, they rated the level of sarcasm in their own responses (Step 3 in Figure 1) by selecting one of the following six options (1 - 6 Likert scale: *not at all*, *mostly not*, *not so much*, *somewhat*, *mostly*, *completely*). We collected ratings in a 6-point scale to obtain labels suitable for different sarcasm modeling tasks (sarcasm classification with binary or multi-class labels and sarcasm quantification). As they were not asked to rate their responses until they freely responded to all situations, participants were not primed to be sarcastic, hence providing natural responses, some of which turned out to be sarcastic due to the influence of the provided context.

Next, we recruited 360 evaluators (180 female, mean age = 35) to rate the level of sarcasm of the responses given by each of the 60 authors of those responses. Six annotators saw the same 32 situations and the responses collected in the response generation process. Using the same 6-point scale, they rated the level of sarcasm of the last response in each situation. As there were 60 response generators in the previous process, the total number of annotators was 360 (6 * 60).

The inter-annotator agreement indicated by

---

[3]The data collection process of part 1 is published at Jang et al. (2023) with analyses unrelated to this work. It provides more information about specific design choices of the stimuli.

[4]We used Prolific and FindingFive for all our data collec-

tion processes.

Table 2: Dataset comparison. A: authors labels, T:third-party labels. -C: without context, +C: with context, Sim.Conv. = Simulated conversations.

	CSC	MUStARD	SC V2	iSarcasmEval
# of sarcastic sentences	2,210 (A) / 2,398 (T)	345	4,693	1,067
# of total sentences	7,040	690	9,386	3,801
# of sent. for training	4,420 (A) / 4,796 (T)	690	9,386	2,134
Average sent. length (-C)	11	12	49	18
Average sent. length (+C)	51	42	-	-
Author labels exist	Y	N	N	Y
Third-party labels exist	Y	Y	Y	N
Is multimodal	N	Y	N	N
Context exists	Y	Y	N	N
Original source domain	Sim. conversations	TV series	Online debates	Social media
Original label type	Multi (1-6)	Binary	Binary	Binary
Annotator agreement	Moderate (Kendall's W 0.56)	Low (Kappa 0.23)	High (Percent agreement 0.80)	N/A

Kendall's W was moderate (0.46).

After successfully collecting 30% of sarcastic responses from part 1 (see Table 3), we collected more data points using the same experimental design but with different situational prompts. We created 40 new situation descriptions and collected responses from 128 native English speakers (64 female, mean age=37) and the sarcasm labels. This time, to increase variability in the generated sentences, participants were asked to imagine being a close friend of the interlocutor for half of the situations and a colleague for the other half. 512 annotators (256 female, mean age=38) rated the level of sarcasm of the responses of each of the response generators. The Kendall's W was 0.56 indicating moderate agreement again.

Table 3: Proportions of sarcastic (S) and non-sarcastic (NS) responses by author labels (A) and third-party labels (T) from the data collection process of CSC (Part 1 and Part 2). The original 1–6 labels were binary-coded for this table.

		Part1	Part2	Total	%
A	NS	1,289	3,537	4,826	69
	S	631	1,579	2,210	31
T	NS	1,307	3,331	4,638	66
	S	613	1,785	2,398	34

The CSC dataset thus has the following structure: context, response, response author id, evaluator id, author label for sarcasm, and third-party label for sarcasm. Out of 7,040 collected sentences, around 30% were judged as sarcastic, which is a higher proportion of sarcasm than is usual (8 - 12%), as reported in prior work (Oraby et al., 2016; Gibbs,

2000). We release the full dataset[5], but we run all subsequent experiments with binary-coded labels (*completely*, *mostly*, *somewhat* to *sarcastic*) and with balanced number of labels after random down-sampling (extracted using the random state of 2) to maximize comparability with the other datasets that are balanced.

3.3 The content of the datasets

Comparing different datasets helps identify the vast spectrum of sarcasm available and further supports the need to test the generalizability of sarcasm detection models on the other datasets for a more robust sarcasm detector system. Table 2 shows points of comparison among the 4 datasets. Apart from these quantitative and structural differences, here we report some qualitative differences (i.e., styles and topics) we find from inspecting each dataset. For CSC and MUStARD, with the contexts mostly being friendly situations, the elements of comedy and friendly mocking are prevalent. The contexts in MUStARD often do not provide enough information to indicate that the following utterance should be sarcastic or not, because more contextual information should be extracted either at the episodic or series level, or from multimodal cues, which often gets lost in the text-only version. Sarcasm Corpus V2 has the most aggressive, critical, and provocative type of sarcasm, which is understandable as the source of the data is online debate forums, from posts on controversial topics such as 'homosexuality', 'abortion', 'gun control', or 'religion'. In contrast, iSarcasmEval has the most instances of self-deprecating type of sarcasm since tweets are often used by individuals to express their own opin-

---
[5]The full data can be found at:
https://github.com/CoPsyN/CSC.

ions on a wide range of everyday topics. Even sarcastic instances that are directed at others in iSarcasmEval exude more distant/cynical attitudes towards the target compared to Sarcasm Corpus V2 (see Example No. 1 of iSarcasmEval in Table 1).

4 Experimental setup

4.1 Models

We experimented with three encoder-only models from Transformers: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021). We limited our models to three widely-used and reliable classification models to focus on cross-data comparisons. We fine-tuned `bert-base-uncased`, `roberta-base`, and `mdeberta-v3-base`. The models were fine-tuned for 2 epochs with a batch size of 64, a learning rate of 5e-5, and a weight decay of 1e-2. We used 5-fold validation and 4 different seeds (1, 11, 21, 31). The fine-tuning was implemented using the Trainer class from the Hugging Face library, and conducted on a NVIDIA A100-PCIE GPU with a total memory of 40GB. All the results reported in Section 5 are an average across all seeds and folds.

4.2 Intra-dataset vs. cross-dataset detection of sarcasm

To test the generalizability of sarcasm detection spanning different datasets, we performed sarcasm detection in two ways: intra-dataset and cross-dataset detection. For the intra-dataset settings, all models fine-tuned on dataset A gave predictions on the hold-out test set of dataset A. For the cross-dataset settings, all models fine-tuned on Dataset A gave predictions on Dataset B. For all of these predictions, we did a 5-fold validation (64% training, 16% validation, 20% test).

5 Results and discussion

Table 4 reports the F-scores of all models in the intra-dataset and cross-dataset conditions.

5.1 Intra-dataset predictions

For all LMs, the best performance is obtained for the Sarcasm Corpus V2 (SC V2), followed by the Conversation Sarcasm Corpus with third-party labels (CSC-T), and the lowest performance on iSarcasmEval. The high performance of SC V2, aside from it being the largest dataset with the most annotators, is also attributed to its source: online forums. In these forums, users can only use text for

communication, potentially leading to a rich concentration of lexical cues associated with sarcasm, such as aggression and negative emotions. These textual cues are more easily identifiable by LMs, enhancing their ability to detect sarcasm. MUStARD fares worse than most other models possibly due to the loss of multimodal cues in text-only settings, which normally would have complemented the long-dependency contexts.

The source of sarcasm labels consistently affects the model performance in the intra-dataset settings. Models fine-tuned on iSarcasmEval, which only has author labels, show the worst performance compared to the others. For CSC also, LMs fine-tuned with third-party labels always perform better than with author labels (**C2**). This observation could suggest that language models may act more as passive observers lacking introspective abilities: Observers, either humans or language models, must rely on external cues and make inferences about an utterance to interpret it since they have no direct access to the complex motivations of the speaker. This aligns with the existence of *sarchasm*, 'sarcasm gone missed', which is a prevalent phenomenon in human communication (Fox Tree et al., 2020).

5.2 Cross-dataset predictions

Overall, all LMs struggle to detect sarcasm on the other datasets proportionately to their performance in intra-dataset settings. Even LMs fine-tuned on SC V2, which showed the highest performance in the intra-dataset predictions, do not generalize nearly as well to the other datasets.

The datasets that result in the most struggle for fine-tuned LMs in generalization are SC V2 (BERT), MUStARD (RoBERTa, DeBERTa), and iSarcasmEval (RoBERTa), based on the average score of each row in Table 4. In contrast, different versions of CSC used in fine-tuning lead to the best performance in stable generalization to the other datasets for all LMs, except for one case (BERT+iSarcasmEval performed the best on SC V2 with an F-score of 0.59).

Findings from previous work employing similar methodologies suggest that, for models to generalize to other datasets effectively, they must first exhibit robust performance on the data used for their fine-tuning (Fortuna et al., 2021) or that datasets should be large (Yin and Zubiaga, 2021; Halevy et al., 2009). But, our results show that the new

Table 4: F-scores of all intra- and cross-dataset predictions. A: with author labels, T: with third-party labels, +CONT: text consisting of both context and utterance. The best fine-tuned LM(s) for each test set marked in bold (columnwise).

Model	fine-tuned on	Intra-dataset	Predicted on — Cross-dataset							
			CSC-A+CONT	CSC-T+CONT	CSC-A	CSC-T	MUS+CONT	MUS	SC V2	iSarcasm
BERT	CSC-A+CONT	0.68	-	-	-	-	0.54	0.56	0.42	0.50
	CSC-T+CONT	0.73	-	-	-	-	0.55	0.57	0.51	**0.53**
	CSC-A	0.67	-	-	-	-	**0.57**	**0.58**	0.39	0.43
	CSC-T	0.70	-	-	-	-	0.56	0.56	0.46	0.47
	MUS+CONT	0.63	0.45	0.46	0.51	0.50	-	-	0.39	0.44
	MUS	0.63	0.47	0.47	**0.53**	**0.52**	-	-	0.40	0.45
	SC V2	**0.77**	0.44	0.44	0.44	0.44	0.39	0.46	-	0.45
	iSarcasm	0.59	**0.48**	**0.48**	0.52	0.51	0.44	0.50	**0.59**	-
RoBERTa	CSC-A+CONT	0.68	-	-	-	-	**0.59**	0.55	0.48	0.52
	CSC-T+CONT	0.72	-	-	-	-	0.57	**0.57**	**0.56**	**0.54**
	CSC-A	0.66	-	-	-	-	0.55	0.56	0.42	0.44
	CSC-T	0.70	-	-	-	-	0.56	**0.57**	0.51	0.51
	MUS+CONT	0.44	0.35	0.35	0.39	0.38	-	-	0.37	0.38
	MUS	0.44	0.35	0.35	0.41	0.40	-	-	0.36	0.40
	SC V2	**0.80**	0.47	0.49	0.52	0.53	0.39	0.49	-	**0.54**
	iSarcasm	0.42	0.36	0.35	0.38	0.39	0.36	0.37	0.44	-
DeBERTa	CSC-A+CONT	0.67	-	-	-	-	**0.55**	**0.57**	0.44	**0.52**
	CSC-T+CONT	0.72	-	-	-	-	**0.55**	0.56	0.53	**0.52**
	CSC-A	0.65	-	-	-	-	0.54	0.55	**0.56**	0.48
	CSC-T	0.69	-	-	-	-	0.53	0.54	0.55	0.50
	MUS+CONT	0.44	0.37	0.37	0.40	0.40	-	-	0.45	0.39
	MUS	0.43	0.35	0.35	0.43	0.41	-	-	0.36	0.40
	SC V2	**0.78**	**0.53**	**0.53**	**0.50**	**0.50**	0.37	0.47	-	0.49
	iSarcasm	0.41	0.34	0.34	0.38	0.37	0.45	0.50	0.35	-

CSC dataset can deal with a broader range of sarcasm despite not being the largest dataset nor yielding the highest accuracy in predictions on its own dataset. Such relative competence could have stemmed from a strong advantage of CSC over other datasets, which is the psycholinguistically-motivated collection and annotation of the data. This suggests that an additional factor in a dataset that contributes to a high generalizability of an LM includes data collection methodology.

The effects of context were tested by fine-tuning the LMs with or without context for CSC and MUS-tARD. For the *with-context* condition, contexts were concatenated with responses for fine-tuning; for the *without-context* condition, only responses were used in fine-tuning. This was done because context is an important aspect in humans' processing of sarcasm but it has been less obvious for LMs (Woodland and Voyer, 2011; Castro et al., 2019; Jaiswal, 2020; Ghosh et al., 2018). Also, testing the effects of context was important for CSC as it has an embedded structure in which different responses are preceded by the same context.

The results showed that LMs fine-tuned with context obtain slightly better results than LMs fine-tuned without it (in intra-dataset predictions), though the improvement was very small (between 0.01 and 0.03). In cross-dataset predictions, the presence of context did not affect the generalizability of the models either. Note that this could be the result of the way context was concatenated to the response or of the number of fine-tuning epochs. For more comprehensive results, future research should address this aspect with targeted experimental manipulations.

The domain of the dataset does not affect the model performance in any obvious way. One might assume that the generalization ability of LMs would benefit from datasets that share a domain (i.e., CSC and MUStARD; both from conversational contexts). This was not the case, as only the LMs fine-tuned on CSC predicted sarcasm on MUStARD well (F-scores 0.53 - 0.59), but not the other way around (F-scores 0.35 - 0.53) except for two cases (BERT+MUS on CSC-A and CSC-T). The LMs that generalized the best on CSC were in fact fine-tuned on SC V2 (for RoBERTa and DeBERTa) or iSarcasmEval or MUStARD (for BERT), most of which come from a different domain than CSC (social media or debate forums). The success of SC V2 may be attributed to the size and annotation quality of the dataset (Yin and Zubiaga, 2021).

Combining all datasets for fine-tuning did not improve model performance. When BERT was fine-tuned on all datasets combined, the average F-score was 0.71, still lower than the two highest performing datasets (SC V2 and CSC-T-CONT). This shows that a mere combination of several datasets with different sizes, styles, and label sources is insufficient in improving the generalizability of sarcasm detection.

6 Post-hoc qualitative analysis

We probe for reasons behind the low generalizability of LMs with a post-hoc analysis. We conducted a quantitative analysis in which we quantitatively analyze and demonstrate how LMs get accustomed to different types of cues for detecting sarcasm when fine-tuned with different datasets (**C3**). We focus on BERT only for this analysis.

Table 5: List of LIWC categories and examples used for Section 6. Original source at: https://mcrc.journalism.wisc.edu/files/2018/04/Manual_LIWC.pdf

Category	Examples
Negations	no, not, never
Positive emotion	love, nice, sweet
Negative emotion	hurt, ugly, nasty
Anxiety	worried, fearful
Anger	hate, kill, annoyed
Sadness	crying, grief, sad
Social processes	mate, talk, they
Family	daughter, dad, aunt
Friend	buddy, neighbor
Cognitive processes	cause, know, ought, think, know, because, should, maybe, always, never
Perceptual processes	look, heard, feeling
Drives	friend, social, win, success, superior, bully, take, prize, benefit, danger, doubt
Religion	altar, church
Swear words	fuck, damn, shit
Online register	btw, lol, thx
Agreement	agree, ok, yes
Nonfluencies	er, hm, umm

Methods. We identified certain patterns of language found in each dataset that may have led to differing results. As a proxy for *styles* or *registers*, we chose 17 semantic and psycholinguistic categories provided by the linguistic analysis tool LIWC (Pennebaker et al., 2015), which quantitatively analyzes text in terms of psychological con-

structs such as emotion, cognition, or perception, among others (See Table 5). LIWC produces the ratio of words belonging to each of these categories per sentence.

Working with one dataset at a time, we took all the sarcastic instances correctly identified as sarcastic 1) *only* by BERT fine-tuned on the same dataset (*own success*) and 2) *only* by BERT fine-tuned on the other datasets (*others' success*). For each of these instances, we obtained LIWC scores for the 17 linguistic categories and averaged them across all the instances. We calculated the difference between the scores of *own success* and *others' success*. The score difference indicates the linguistic properties that facilitate the detection of sarcasm that are uniquely present in one dataset (and not in the others).
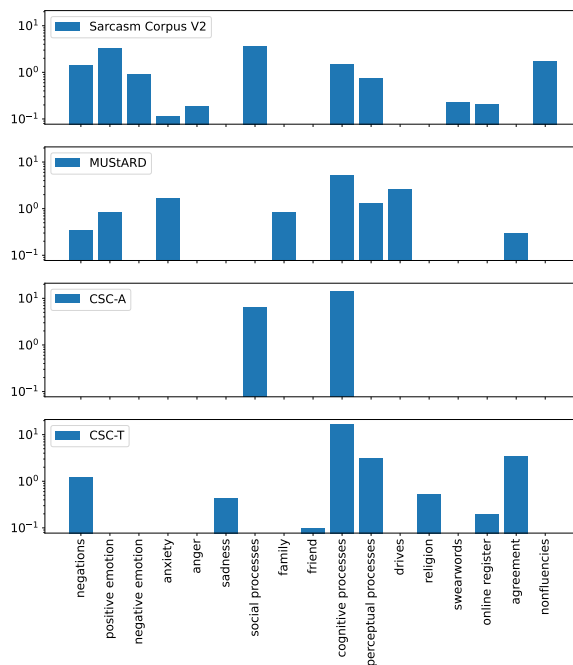

Figure 2: Post-hoc analysis to quantify diverging features found in different datasets of sarcasm. The height of each bar represents the log transformed difference between the score for each linguistic property found in sentences uniquely detected by LMs fine-tuned on the same data and those uniquely detected by LMs fine-tuned on the other data. The full description of the categories is provided in Table 5.

Results. Figure 2 reports the unique features that help BERT detect sarcasm in each dataset. A major characteristic of **Sarcasm Corpus V2** is that it contains many words related to negative emotions such as anger, and words related to social issues, swear-words, online-style words, and disfluencies. This matches our preliminary description of the dataset

described in Section 3.3 that Sarcasm Corpus V2 contains the most aggressive and critical types of sarcasm of out of all the datasets. **MUStARD** contains a lot of words related to family and drives (i.e., achievement, risk, reward, etc.). We suspect that this phenomenon may be influenced by the TV series The Big Bang Theory, where the characters often converse on topics related to achievement. **CSC** shows a lot of words related to agreement (i.e., Ok, yes, etc.) possibly because the speakers were instructed to talk to a close friend or colleague. The analysis also reveals a significant presence of words associated with religion, predominantly due to the use of terms like 'god' or 'Jesus' in expressions such as 'oh my god' or 'Jesus Christ'. This trend might be attributed to numerous situational prompts where the interlocutor's behavior is portrayed as silly, eliciting strong emotional responses. The results for **iSarcasmEval** are not reported because there were no instances for which fine-tuning on iSarcasmEval only and not on the other datasets produced correct predictions.

The post-hoc analysis of different datasets and model performance shows that sarcasm does indeed come in different shapes and domains, some serious and genuinely meant to inflict verbal pain to strangers, and some humorous and friendly used among well-meaning acquaintances, which is also supported by prior psycholinguistic work (Bowes and Katz, 2011; Colston, 1997; Dews et al., 1995; Frenda et al., 2022; Gibbs, 2000; Pexman and Olineck, 2002).

7 Conclusion

We conducted intra- and cross-dataset sarcasm detection experiments to examine the robustness of sarcasm detection models. We introduced a new dataset, which we compared with other existing datasets of varying characteristics. Several characteristics regarding sarcasm in each dataset, such as label source (authors vs. third-party), domain (social media/online vs. offline conversations/dialogues), style (aggression vs. harmless mocking) were used as points of comparison. All LMs gave better predictions on the same dataset they were fine-tuned on and showed much lower prediction performance on the other datasets (**C1**). Still, LMs fine-tuned on our new dataset CSC generalized the best to the other datasets. This was the case even when the domain of sarcasm in the target dataset was different from that of CSC.

LMs performed better when the ground-truth labels were from third-party annotators rather than authors themselves (**C2**). A post-hoc analysis and a closer look at each dataset supported our assumption that such low cross-dataset predictions may be attributed to sarcasm coming in various styles, intent, and shapes (**C3**). Therefore, future research should take the vast scope of sarcasm into consideration rather than only focusing on a narrow definition of sarcasm as 'the utterance of the opposite of the intended meaning' or as 'a figure of speech intended to be hurtful, insensitive, and critical'.

Limitations

We limited the number of datasets and language models to focus on the performance comparison of language models fine-tuned on different datasets. Other publicly available sarcasm datasets not used in this work include SAD (Painter et al., 2022), SARC (Ptáček et al., 2014), and MUStARD++ (Ray et al., 2022). The language models mentioned in this paper exclude generative models or dialogue-focused models. Experiments employing other models with more parameters, other datasets, and with different hyperparameters (i.e., different number of epochs, etc.) are left to future work.

Ethics Statement

We generally see little ethical issue related to this work. All our experiments involving human participants were conducted on a voluntary basis with a fair compensation suggested by the recruitment platform Prolific and participants were informed on how the data will be used. All our modeling experiments were conducted with open-source libraries, which received due citations. The experiment is in line with the ethical regulations of the University of Konstanz (IRB number 05/2021) . But we acknowledge that at times, sarcasm in itself can be a sensitive topic including offensive language and content depending on the circumstances.

Acknowledgements

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452, Portorož, Slovenia. European Language Resources Association (ELRA).

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814, Seattle, United States. Association for Computational Linguistics.

Muhammad Abulaish and Ashraf Kamal. 2018. Self-Deprecating Sarcasm Detection: An Amalgamation of Rule-Based and Machine Learning Approach. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 574–579. IEEE. Event-place: Santiago.

Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. 2014. Modelling sarcasm in twitter, a novel approach. In *proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 50–58.

Alexandru-Costin Băroiu and Ștefan Trăușan-Matu. 2022. Automatic sarcasm detection: Systematic literature review. *Information*, 13(8):399.

Andrea Bowes and Albert Katz. 2011. When Sarcasm Stings. *Discourse Processes*, 48(4):215–236.

Penelope Brown and Stephen Levinson. 1978. Universals in language usage: Politeness phenomena. In Esther Goody, editor, *Questions and Politeness: Strategies in Social Interaction*, pages 56–310. Cambridge University Press, Cambridge.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multimodal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Herbert L Colston. 1997. Salting a wound or sugaring a pill: The pragmatic functions of ironic criticism. *Discourse processes*, 23(1):25–45.

Sourav Das, Soumitra Ghosh, Anup Kumar Kolya, and Asif Ekbal. 2023. Un paralleled sarcasm: a framework of parallel deep lstms with cross activation functions towards detection and generation of sarcastic statements. *Language Resources and Evaluation*, 57(2):765–802.

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.

Shelly Dews, Joan Kaplan, and Ellen Winner. 1995. Why not say it directly? The social functions of irony. *Discourse Processes*, 19(3):347–367.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Lrec*, pages 392–398. Citeseer.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing Management*, 58(3):102524.

Jean E. Fox Tree, J. Trevor D'Arcey, Alicia A. Hammond, and Alina S. Larson. 2020. The sarchasm: Sarcasm production and identification in spontaneous conversation. *Discourse Processes*, 57(5-6):507–533.

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 and Tony Veale. 2017. Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491. Association for Computational Linguistics. Event-place: Copenhagen, Denmark.

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

Raymond W. Gibbs. 2000. Irony in Talk Among Friends. *Metaphor and Symbol*, 15(1-2):5–27.

Sam Glucksberg. 1995. Commentary on nonliteral language: Processing and use. *Metaphor and Symbol*, 10(1):47–57.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE intelligent systems*, 24(2):8–12.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

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.

Julia Jorgensen. 1996. The functions of sarcastic irony in speech. *Journal of Pragmatics*, 26(5):613–634.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic Sarcasm Detection: A Survey. *ACM Computing Surveys*, 50(5):1–22.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing Context Incongruity for Sarcasm Detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762. Association for Computational Linguistics. Event-place: Beijing, China.

Thomas R Keenan and Kathleen Quigley. 1999. Do young children use echoic information in their comprehension of sarcastic speech? a test of echoic mention theory. *British Journal of Developmental Psychology*, 17(1):83–96.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A Large Self-Annotated Corpus for Sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, page 6. European Language Resources Association (ELRA).

Roger J Kreuz and Sam Glucksberg. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of experimental psychology: General*, 118(4):374.

John S. Leggitt and Raymond W. Gibbs. 2000. Emotional Reactions to Verbal Irony. *Discourse Processes*, 29(1):1–24.

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*.

Jacqueline K. Matthews, Jeffrey T. Hancock, and Philip J. Dunham. 2006. The Roles of Politeness and Humor in the Asymmetry of Affect in Verbal Irony. *Discourse Processes*, 41(1):3–24.

Silviu Oprea and Walid Magdy. 2019. Exploring Author Context for Detecting Intended vs Perceived Sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2859, Florence, Italy. Association for Computational Linguistics.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A Dataset of Intended Sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Silviu Oprea, Steven Wilson, and Walid Magdy. 2021. Chandler: An explainable sarcastic response generator. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 339–349, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41, Los Angeles. Association for Computational Linguistics.

Jordan Painter, Helen Treharne, and Diptesh Kanojia. 2022. Utilizing weak supervision to create S3D: A sarcasm annotated dataset. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 197–206, Abu Dhabi, UAE. Association for Computational Linguistics.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392, Online. Association for Computational Linguistics.

James Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. Technical report, University of Texas at Austin.

Penny M. Pexman and Kara M. Olineck. 2002. Does Sarcasm Always Sting? Investigating the Impact of Ironic Insults and Ironic Compliments. *Discourse Processes*, 33(3):199–217.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 213–223.

Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. 2022. A multimodal corpus for emotion recognition in sarcasm. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6992–7003, Marseille, France. European Language Resources Association.

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. Association for Computational Linguistics.

Dan Sperber and Deirdre Wilson. 1981. Irony and the use-mention distinction.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Citeseer.

Martin Sykora, Suzanne Elayan, and Thomas W Jackson. 2020. A qualitative analysis of sarcasm, irony and related# hashtags on twitter. *Big Data & Society*, 7(2):2053951720972735.

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*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey. European Language Resources Association (ELRA).

Jennifer Woodland and Daniel Voyer. 2011. Context and Intonation in the Perception of Sarcasm. *Metaphor and Symbol*, 26(3):227–239.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Yazhou Zhang, Yaochen Liu, Qiuchi Li, Prayag Tiwari, Benyou Wang, Yuhua Li, Hari Mohan Pandey, Peng Zhang, and Dawei Song. 2021. Cfn: A complex-valued fuzzy network for sarcasm detection in conversations. *IEEE Transactions on Fuzzy Systems*, 29(12):3696–3710.