Sarcasm Detection Models are *So* Easy to Improve! /s

Diana Abagyan
dabagyan@uw.edu

Libby Merchant
emercha@uw.edu

Emma McKibbin
echm@uw.edu

Jade Phoreman
jphore@uw.edu

Catherine Ball
catball@uw.edu

Abstract

In the present study, we build a simple sarcasm detection system using BERT, fine-tuned on the SARC dataset, then attempt to improve this model by leveraging context and other pre-trained embedding models. As our study continues, we plan to adapt our model to detect sarcasm in neurodivergent communities, which may shed light on how tone indicator usage varies by community.

1 Introduction

Sarcasm detection is a notoriously difficult task, both for the detector and for the task developer. The Self-Annotated Reddit Corpus (SARC) (Khodak et al., 2018) leverages the use of explicit tone indicators in online communities to more reliably annotate their task data. These tone indicators are popular in neurodiverse communities, in part because neurodivergent people are less confident in their interpretation of non-literal language online. We train our initial system on SARC, which pulls data from general populations on Reddit. Then we adapt our system on a self-collected task built from posts in neurodiverse Reddit communities.

Previous studies make use of a variety of models for sarcasm detection, such as SVMs (Tungthamthiti et al., 2016), CNNs (Hazarika et al., 2018), and ensemble approaches adding LSTMs and MLPs (Lemmens et al., 2020). In general, the literature shows a shift from using GloVe to using BERT (Devlin et al., 2019) for embeddings, so we begin with an initial system using BERT as our classifier. We then revise our system in two ways: first, by augmenting our input with the previous comment as context; second, by using RoBERTa (Liu et al., 2019) as our embedding model in place of BERT. We find that the model with both of these revisions performs best on our development set.

2 Related Work

Automatic sarcasm detection tasks have historically posed serious challenges as they are strongly dependent upon context. A variety of model architectures have been proposed, such as Support Vector Machines (Tungthamthiti et al., 2016), and ensemble approaches (Lemmens et al., 2020). However, performance varies depending on data source and conversational context provided. Baruah et al. (2020) shows that contextual improvements for a sarcasm detection system on Twitter data may not improve a system processing Reddit data. Models that perform well on Reddit data, such as CASCADE in Hazarika et al. (2018), may make use of non-linguistic data, like user behavior, to better predict sarcasm. However, strategies like these not only require user data that may not be published with every dataset but also run the risk of introducing unethical bias into the system.

Further, the task of sarcasm detection has largely relied on labeled datasets created by third-person annotators. But this task is difficult not just for models but for humans, as well. As Khodak et al. (2018) reports, when human evaluators were presented with two statements and asked to identify which of the two was sarcastic, the average human F1-score was 0.816. Therefore, when developing sarcasm detection datasets, labeling data as sarcastic or non-sarcastic is non-trivial, as it relies strongly upon subjective human judgement.

Of note, neurodivergent (ND) individuals have been shown to process sarcasm in different ways from those who are typically developing: Zalla et al. (2014) indicated that Autism Spectrum Disorder (ASD) individuals are less likely to respond to social stereotypes as a marker of sarcasm, while Ludlow et al. (2017) suggested that children with Attention-Deficit Hyperactivity Disorder (ADHD) are less likely to comprehend paradoxical sarcasm in particular.

Previous studies have not reached a consensus as to whether ND individuals show a deficit in detecting ironic language when overall language ability is controlled for (Kalandadze et al., 2018). In an eye-tracking study on ASD and typically-developing participants, Au-Yeung et al. (2015) found similar accuracy in sarcasm detection between the two groups; however, the ASD group took more time to read and process the statements, suggesting that these individuals have lower confidence in their ability to detect irony. For some in these groups, the task of sarcasm detection poses its own challenge.

In response to this, tone indicators have emerged in neurodiverse communities online as a way to compensate for the unique difficulties of communication in online spaces (Febiana Christanti et al., 2022). Online text does not offer the paralinguistic cues such as prosody, facial expressions, and body language that help disambiguate sarcasm in face-to-face interactions.

Additionally, this linguistic innovation has been explored as a way to circumnavigate the challenges of developing labeled corpora for automated sarcasm detection. The Self-Annotated Reddit Corpus (SARC) put forth by Khodak et al. (2018) proposed a way to extract self-annotated sarcasm datasets through tone indicators in social media posts. This is particularly salient for the purpose of disambiguating ironic language in neurodiverse spaces, due to the differing reactions to ironic language within these groups.

3 Task Description

We develop a sarcasm detection model first on our primary task, which uses the SARC datset. Then, we adapt our model to our self-collected adaptation task, which differs from SARC primarily in the population producing the self-annotated posts. In SARC, the population is general; in our dataset, the population is restricted to communities with high concentrations of ND members.

3.1 Primary Task

Our primary task is sarcasm detection using a model trained on the SARC dataset.[1] This task is characterized by the following dimensions:

- **Affect type**: emotion (sarcasm)
- **Recognition type**: classification
- **Genre**: Reddit posts
- **Target**: N/A
- **Modality**: text
- **Language**: English

The SARC dataset was collected from Reddit posts and comprises 533M comments total, with 1.3M sarcastic comments. All comments include data about the author, topic, and context. The data is self-annotated in the sense that users labeled their own sarcastic comments using the tone indicator */s*.

We utilize the balanced portion of the dataset during model development. The sizes of each partition used in training, development, and testing are described in Table 1. Note that we created a development set from 10% of the provided balanced training set.

Partition	Count
Original Balanced Training	257,082
Training	231,374
Development	25,708
Balanced Testing	64,666

Table 1: The size of each partition from the SARC dataset used in our study. We train on "Training without Dev," do intermediate evaluation on "Development," and finally evaluate on "Balanced Testing."

In the SARC data, there is some level of noise associated with self-annotation using tone indicators; some sarcastic statements may not be labeled as such, and those that are labeled may only be as such because they are especially ambiguous (Khodak et al., 2018). The first source of noise can be mitigated in our own data collection by targeting only those posts labeled as serious (represented by */srs*), but this exacerbates the second source of noise.

3.2 Adaptation Task

Our adaptation task is another sarcasm detection task, using a model fine-tuned with data we've collected specially for this study. The data is sourced from online communities we suppose have significantly higher proportions of ND users. We collected text posts from subreddits such as r/Neurodivergent, r/neurodiversity, and r/autism, where self-labeling sarcasm with the */s* tone indicator is somewhat common. There does not seem to be any literature on using computational models to

[1] SARC data: https://nlp.cs.princeton.edu/old/SARC/2.0/

perform sarcasm detection among ND populations specifically, so instead we refer to adjacent literature for guidance, such as Au-Yeung et al. (2015) and Febiana Christanti et al. (2022).

3.2.1 Data Collection

Our adaptation task necessitated the creation of a new dataset primarily authored and labeled by ND netizens. Reddit offers a convenient way to narrow the search for this data due to its prominence of neurodiversity-related forums, or *subreddits*. To target users belonging to these communities, we limited our data sources to relevant subreddits, listed in Appendix A.

The final dataset merged data points from an initial scrape and data retrieved from the Pushshift Reddit Dataset (16.8M instances). There was no overlap between the two given the different time frames in which they were collected.

4 System Overview

Initial System The baseline system architecture of the present study was adapted from (Baruah et al., 2020) and comprises a pre-trained BERT model with a classification layer. For this, we used BertForSequenceClassification as accessed through the transformers library. We then fine-tuned this model for sarcasm detection before performing classification on our primary and adaptation tasks.

RoBERTa Liu et al. (2019) reproduced the original BERT experiment, paying extra attention to the effects of the model's hyperparameters and the training task. Most notably, their model, RoBERTa, underwent significantly more pre-training than BERT, resulting in a model that outperforms BERT on at least one sentiment analysis task. For this reason, we test whether using RoBERTa in place of BERT for embeddings and classification can improve our system's performance. We use the RobertaForSequenceClassification class in HuggingFace, which also adds a linear layer for classification after the RoBERTa model.

Adding Context Baruah et al. (2020) experiment with 5 forms of including context- not at all, including the last utterance from the dialogue, the last two utterances, the last three utterances, or the last three. They found that for Reddit data, using only the response performed the best, while for Twitter, having a context window that includes only the last utterance was the best-performing. We experiment with including the last post and context, and adding no context at all.

Following Baruah et al. (2020), the context is concatenated with the response in reverse order as they appear in the discourse, separated by a special token.

Combined System Lastly, as our two revisions are compatible for simultaneous use, we trained a system that uses RoBERTa *and* considers previous context.

4.1 Evaluation Methodology

We evaluate all of our systems using F1-score, as done in Khodak et al. (2018), as it considers both precision and recall. A true positive occurs when the correct class labels of "sarcastic" (1) and "non-sarcastic" (0) matches the predicted class label given by the model.

5 Approach

For our primary task, we first partitioned and reformatted the data so that it would be easily accessible during training. Then, we trained each system on a personal GPU and evaluated the systems on our development set using Patas.

5.1 Pre-processing

The balanced training data is provided in .csv format. Each line includes the comment ID for the original post, IDs for any intervening comments, the two IDs of the sarcastic and non-sarcastic responses, and finally the label (0 or 1) for the respective responses. The full comment texts are available in a single JSON file, but because the full dataset comprises 533M comments, accessing this JSON file even once is time- and resource-intensive.

To combat this overhead, we loaded the full file once and printed smaller JSONs for the training and development sets, as described in Table 1. In the final JSONs, each line represents a single sarcastic or non-sarcastic response and contains the text for all previous comments and the response, as well as all the post IDs and the true label. An example is shown in Appendix B.

We use the native BERT/RoBERTa tokenizers, which prepend a [CLS] token, add a [SEP] token at the end of the text, and pad the input out to the maximum model length, 512. If the input is longer than the maximum model length, the remainder is truncated. We use a batch size of 8, so padding the input is necessary.

5.2 Model Training

Each model was trained on an NVIDIA GeForce RTX 2080 Ti GPU, and training lasted 4 hours. According to the suggestions of Devlin et al. (2019), we trained for 4 epochs. Training for many epochs could lead to catastrophic forgetting. To further mitigate this, we used the lowest suggested learning rate of 2e-5, as suggested in Sun et al. (2020). We keep the model checkpoint saved from the epoch with the best F1 score on the dev set.

5.3 Model Evaluation

We tested the model on a dev set we created using 10% of the SARC balanced training data, randomly sampled. At the same time, we tracked evaluation metrics such as F1-score, model loss, and predicted labels and gold standard labels. We later used these labels to analyze the breakdown of results into true/false, negative/positive.

Evaluating all 4 of our models took 30 minutes and 31 seconds on the Linguistic Department's computing cluster Patas. We requested and were allocated a GPU, though it is difficult to tell retroactively which node we were allocated and how much of our job time was spent on the GPU.

6 Results

The F1-scores for each of our 4 models are shown in Table 2 alongside random and embedding baselines. The best performing model was our combination model, which uses RoBERTa for word embeddings and provides the previous comment for context. The F1-score improvement from our initial system is +0.011.

Model	F1
RANDOM (BASELINE)	0.500
BERT	0.725
ROBERTA	0.729
BERT + CONTEXT	0.727
ROBERTA + CONTEXT	**0.736**

Table 2: Each model's F1-score on the dev set. The random baseline score is as reported in Khodak et al. (2018). Note that these two models were evaluated on SARC's balanced test set, whereas our models are evaluated on our dev set.

7 Discussion

F1-scores for our primary task show improvement over the baseline as well as tentative improvement over model performance on the Reddit dataset in (Baruah et al., 2020). False positives for our baseline system are slightly more numerous than false negatives in the dataset, at 3769 and 3401, respectively.

	Model Label		
	Sarcastic	Non-sarcastic	Total
True	9453	9085	18,538
False	3769	3401	7170
Total	13,222	12,486	

Table 3: Confusion matrix of true and false prediction results on development dataset (N=25,708).

Because the dataset is perfectly balanced, we would expect to see an even number of false positives and false negatives, suggesting a systematic error. One cause of this may be simple noise from the dataset: a lack of the /s tone indicator does not guarantee that an utterance is serious. For example, analysis of false positives found in the SARC dataset returns entirely un-serious phrases like the following:

> "It's obviously tracks from a giant water tractor, farming for giant arctic sea prawn!" [1]

> "Because OBAMMA IS A SERKIT MUSLIN." [75]

Table 3 displays a confusion matrix for outputs of the baseline system. Similarly, a comparison of techniques for model enhancement can be found in Table 4. While the Context and RoBERTa models individually perform at comparable levels in terms of F1 scores, further breakdown shows that the RoBERTa may demonstrate a tendency to over-classify instances as sarcastic (n=13,676) compared to the baseline (n=13,222) and Context-only (n=12,878).

Of note, the combination of the two models resulted in identical numbers to the RoBERTA-only model for false negatives (FN) and true positives (TP). It appears that the addition of context then "converted" some of the false positives (FP) into true negatives (TN), resulting in an improved combined score. A similar phenomenon can be observed when comparing the Context-only model to the baseline; the addition of context led to a decrease in false positives (-249) and an increase in true negatives (+249) while other metrics were relatively consistent.

	Context	RoBERTa	Combined
TP	9358	9675	9675
TN	9334	8853	9084
FP	3520	4001	3770
FN	3496	3179	3179

Table 4: Comparison of true and false prediction results across various models.

It is unclear whether the addition of context contributes directly to an increase in true negatives, and if so, why context plays such an important role in identifying negative instances. One possible explanation is that context-agnostic models rely heavily on lexical cues that could signal sarcastic language in the absence of other context.

8 Ethical Considerations

Ethical concerns came into play at several points in the present study. A main concern is data rights and compliance with IP law. As of the time that data scraping is performed, Reddit does not permit publication of their data without express approval. We addressed this by omitting the data that was scraped by the authors from our public repository. Another, perhaps greater, concern is that it was not possible to obtain consent from every Reddit user whose post was featured in the dataset. In response to this, the only assurance we can provide is that the work conducted in the present study will not monetize this data or attempt to publicize any data that has not already been publicized. Further, PII such as usernames were removed from the data.

9 Conclusion

For future work, we plan to address the sources of noise mentioned throughout this paper: (1) sarcastic utterances that are not marked with tone indicators, (2) overabundance of sarcastic utterances that are more confusing or ambiguous than usual, and (3) sarcastic utterances posted by non-ND netizens.

References

Sheena K. Au-Yeung, Johanna K. Kaakinen, Simon P. Liversedge, and Valerie Benson. 2015. Processing of written irony in autism spectrum disorder: An eye-movement study. *Autism Research*, 8(6):749–760.

Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. 2020. Context-aware sarcasm detection using BERT. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 83–87, Online. 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.

Maria Febiana Christanti, Puri Bestari Mardani, and Khansa Ayu Fadhila. 2022. Analysing the meaning of tone indicators by neurodivergent community in twitter. *International Journal of Social Science Research and Review*, 5(1):5–15.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. CASCADE: Contextual sarcasm detection in online discussion forums. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tamar Kalandadze, Courtenay Norbury, Terje Nærland, and Kari-Anne B Næss. 2018. Figurative language comprehension in individuals with autism spectrum disorder: A meta-analytic review. *Autism*, 22(2):99–117.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jens Lemmens, Ben Burtenshaw, Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. Sarcasm detection using an ensemble approach. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 264–269, Online. Association for Computational Linguistics.

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 e-prints*, arXiv:1907.11692.

Amanda K Ludlow, Eleanor Chadwick, Alice Morey, Rebecca Edwards, and Roberto Gutierrez. 2017. An exploration of sarcasm detection in children with attention deficit hyperactivity disorder. *Journal of Communication Disorders*, 70:25–34.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification? *Preprint*, arXiv:1905.05583.

Piyoros Tungthamthiti, Kiyoaki Shirai, and Masnizah Mohd. 2016. Recognition of sarcasm in microblogging based on sentiment analysis and coherence identification. *Journal of Natural Language Processing*, 23(5):383–405.

Tiziana Zalla, Frederique Amsellem, Pauline Chaste, Francesca Ervas, Marion Leboyer, and Maud Champagne-Lavau. 2014. Individuals with autism spectrum disorders do not use social stereotypes in irony comprehension. *PloS one*, 9(4):e95568.

A Subreddits Scraped

Below is a list of the subreddits from which we collected our adaptation task data. We selected subreddits which we felt had primarily a neurodivergent userbase.

- r/ADHD
- r/adhdwomen
- r/aspergirls
- r/AutismTranslated
- r/autismmemes
- r/AutisticPride
- r/Autism
- r/AutisticAdults
- r/autisminwomen
- r/neurodivergent
- r/NeurodivergentLGBTQ

B Data Input Examples

{"posts": ["Which one would you buy?"], "post_ids": ["yae7e"], "context_size": 1, "response_id": "c5u2m60", "response": "The $594.70+37.50$ version, because it obviously has more value.", "label": "1"}

{"posts": ["Which one would you buy?"], "post_ids": ["yae7e"], "context_size": 1, "response_id": "c5tsz3h", "response": "I'm surprised the second one doesn't have $640 shipping fees.", "label": "0"}