Sarcasm Detection

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

This is where the abstract will go, but we will write it last.

1 Introduction

Automatic sarcasm detection tasks have historically posed serious challenges as they are strongly dependent upon context. A variety of model architectures have been proposed, including: Support Vector Machines (Tungthamthiti et al., 2016), ensemble approaches (Lemmens et al., 2020), TODO: Discuss different approaches to sarcasm detection task

Further, the task of sarcasm detection has largely relied on labeled datasets created by third-person annotations. However, such datasets rely strongly upon human judgement and are subject to difference in opinion. 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.

2 Task Description

2.1 Primary Task

In the present study, we will endeavor to expand the task of sarcasm detection to neurodiverse groups. First, our primary task is sarcasm detection using a model trained on the SARC dataset, found at (link) and 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 an unbalanced dataset with 1.3 million sarcastic statement, each with author,

topic, and context. The data is self-annotated in the sense that some users labeled sarcastic comments using the tone indicator */s*.

Notably, 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 by targeting only those posts labeled as serious (represented by */srs*), but this exacerbates the second source of noise. We will discuss these mitigation techniques and others later in the paper.

2.2 Adaptation Task

Our adaptation task is sarcasm detection using a model trained on a dataset we will collect, from a significantly higher proportion of neurodivergent (ND) users. We will collect text posts from subreddits such as r/Neurodivergent, r/neurodiversity, and r/autism, many of which will be self-labeled as sarcastic with the /s tone indicator. There does not seem to be any literature on using computational models to 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 System Overview

The 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 two datasets: one collected from ND utterances and one not.

Our evaluation methodology involved comparing the correct class labels of "sarcasm" and "non-sarcasm" to the predicted class labels given by the model. Using the evaluation metrics as used in (Khodak et al., 2018), we evaluated the model's performance in classifying instances.

4 Approach

4.1 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 *subreddits*: themed forums used predominantly by members of specific communities. As such, data sources were limited to the following subreddits:

- r/neurodiversity
- LIST ALL THE SUBREDDITS USED

The final dataset merged data points scraped by the authors (idk how many instances) 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. A summary of the combined dataset, which we will refer to as INSERT COOL NAME, can be found in figure 2. TODO: CREATE TABLE

4.2 Preprocessing

4.3 Model Building

We fine-tuned our BERT model on 90% of the SARC training data, setting aside the other 10% for our development set. An overview of the training and dev data for the initial task are summarized in figure 1. TODO: CREATE TABLE

4.4 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, predicted labels and gold standard labels, and breakdown of results into true/false, negative/positive.

5 Results

The F1 score resulting from training was 0.909 and model loss of 0.559. The test F1 score was 0.7099, with a loss of 1.4217.

6 Discussion

F1 scores for our primary task show tentative improvement over model performance on the Reddit dataset in (Baruah et al., 2020). False positives (predicted=1, true=0) are slightly more numerous than false negatives in the dataset, at 4233 and 3451, respectively.

Because the dataset is perfectly balance, 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]

7 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.

8 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.

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.

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.

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.

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.