Sarcasm Detection using Context Separators in Online Discourse

Tanvi Dadu * and **Kartikey Pant** *
Netaji Subhas Institute of Technology, New Delhi, India
International Institute of Information Technology, Hyderabad, India
tanvid.co.16@nsit.net.in
kartikey.pant@research.iiit.ac.in

Abstract

Sarcasm is an intricate form of speech, where meaning is conveyed implicitly. Being a convoluted form of expression, detecting sarcasm is an assiduous problem. The difficulty in recognition of sarcasm has many pitfalls, including misunderstandings in everyday communications, which leads us to an increasing focus on automated sarcasm detection. In the second edition of the Figurative Language Processing (FigLang 2020) workshop, the shared task of sarcasm detection released two datasets, containing responses along with their context sampled from Twitter and Reddit.

In this work, we use $RoBERTa_{large}$ to detect sarcasm in both the datasets. We further assert the importance of context in improving the performance of contextual word embedding based models by using three different types of inputs - *Response-only*, *Context-Response*, and *Context-Response (Separated)*. We show that our proposed architecture performs competitively for both the datasets. We also show that the addition of a separation token between context and target response results in an improvement of 5.13% in the *F1-score* in the Reddit dataset.

1 Introduction

Sarcasm is a sophisticated form of speech, in which the surface meaning differs from the implied sense. This form of expression implicitly conveys the message making it hard to detect sarcasm in a statement. Since speech in sarcasm is dependent in context, it is tough to resolve the speaker's intentions unless given insights into the circumstances of the sarcastic response. These insights or contextual information may include the speaker of the response, the listener of the response, and how its content relates to the preceding discourse.

Recognizing sarcasm is critical for understanding the actual sentiment and meaning of the discourse. The difficulty in the recognition of sarcasm causes misunderstandings in everyday communication. This difficulty also poses problems to many natural language processing systems, including summarization systems and dialogue systems. Therefore, it is essential to develop automated sarcasm detectors to help understand the implicit meaning of a sarcastic response.

The sarcasm detection shared task held at the second edition of the Figurative Language Processing (FigLang 2020) workshop proposes two datasets from different social media discourse platforms for evaluation. The first dataset contains user conversations from Twitter, while the second dataset contains Reddit conversation threads. Both datasets contain contextual information in the form of posts of the previous dialogue turns. Their primary aim is to understand the importance of conversational contextual information in improving the detection of sarcasm in a given response.

In this work, we explore the use of contextualized word embeddings for detecting sarcasm in the responses sampled from Reddit as well as Twitter. We outline the effect of adding contextual information, from previous dialogue turns, to the response, for both the datasets. We further explore the importance of separation tokens, differentiating discourse context from the target response while detecting sarcasm.

2 Related Works

Davidov et al. (2010) approached the task of sarcasm detection in a semi-supervised setting, investigating their algorithm in two different forms of text, tweets from Twitter, and product reviews from Amazon. Subsequently, González-Ibáñez et al. (2011) explored this task in a supervised set-

* Both authors contributed equally to the work.

Proceedings of the Second Workshop on Figurative Language Processing, pages 51–55

https://doi.org/10.18653/v1/P17

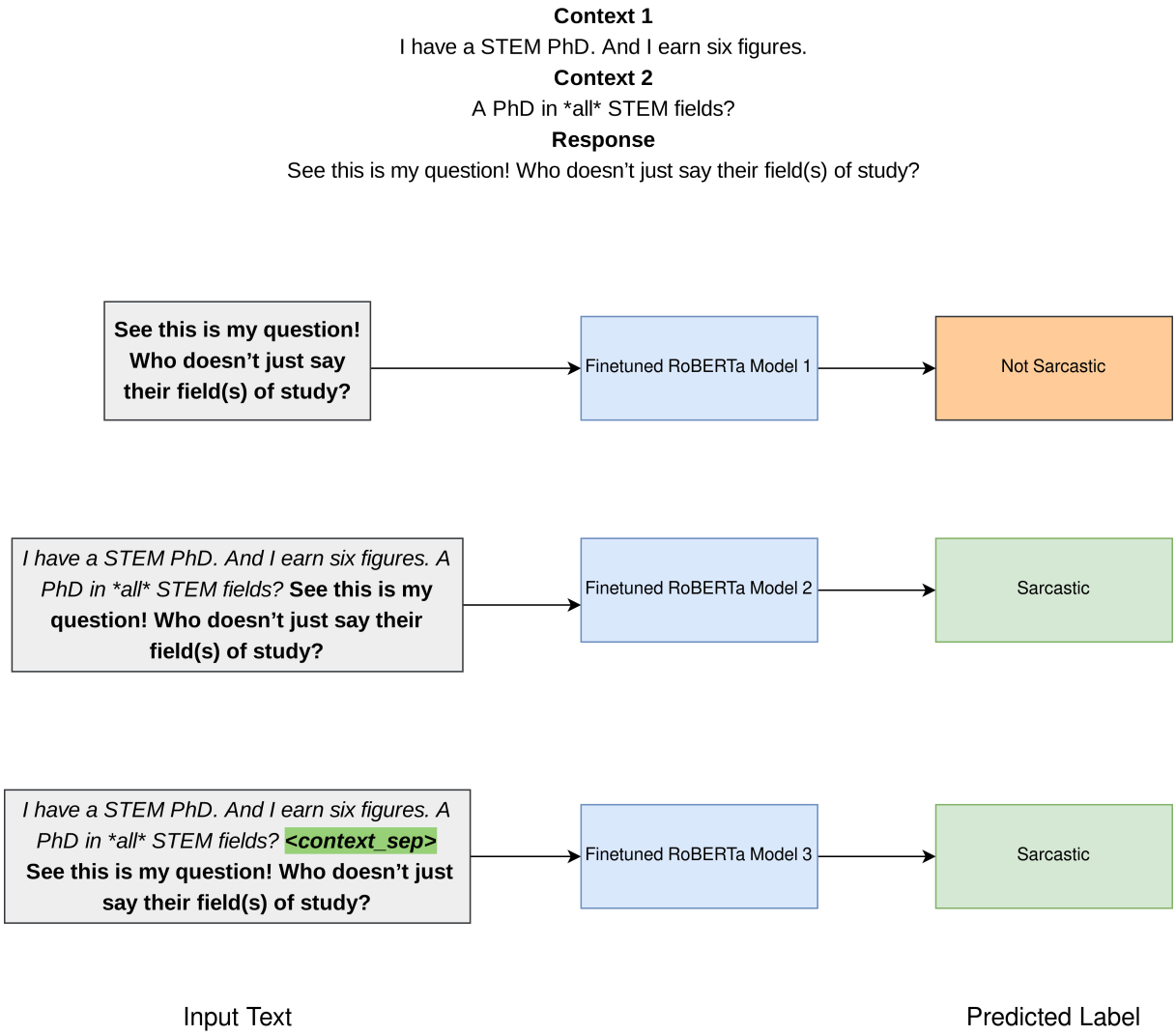

Figure 1: Architecture of our proposed approach.

ting, using SVM and logistic regression. They also release a dataset of 900 tweets to the public domain, entailing tweets containing sarcastic, positive, or negative content. Moreover, there has been significant work done to detect sarcasm in a multi-modal setting, primarily including visual cues. While Schifanella et al. (2016) proposed a multi-modal methodology to exploit features from both text and image, Cai et al. (2019) investigated instilling attribute information from social media posts to propose a model leveraging hierarchical fusion.

Ghosh et al. (2018) investigated the use of Long Short-Term Memory (LSTM) networks with sentence-level attention for modeling both the responses under consideration and the conversation context preceding it. They execute their models on responses sampled from two social media platforms, Twitter and Reddit. They concluded that contextual information is vital for detecting sarcasm by showing that the LSTM network modeling the context and the response outperforms the LSTM network that models only the target response. This observation is in sync with other similar tasks like irony detection, as highlighted by Wallace et al. (2014), which claimed that human annotators consistently rely on contextual information to make judgments.

The use of other turns of dialogue as contextual information has been explored well in previous works. Bamman and Smith (2015) investigated the use of "author and addressee" features apart from the conversation context, but observed only a minimal impact of modeling conversation context. Oraby et al. (2017) investigated using "pre" and "post" messages from debate forums and Twitter

conversations to identify whether rhetorical questions are used sarcastically or not. They observe that using the "post" message as context improved the *F1-score* for the sarcastic class. Joshi et al. (2016) showed that sequence labeling algorithms outperform traditional classical statistical methods, obtaining a gain of 4% in *F1-score* on their proposed dataset.

The use of pre-trained contextual word representations has been explored in multiple tasks in NLP, including text classification. Recently released models such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2020), exploit the use of pre-training and bidirectional transformers to enable efficient solutions obtaining state-of-the-art performance. Pre-trained embeddings significantly outperform the previous state-of-the-art in similar problems such as humor detection (Weller and Seppi, 2019), and subjectivity detection (Pant et al., 2020).

3 Approach

This section outlines the approach used to detect sarcasm in the Reddit and Twitter datasets. Our approach utilizes contextualized word embeddings with three different inputs, as explained in the following paragraphs.

The use of pretrained contextualized word embeddings has been applied to achieve state-of-the-art results for various downstream tasks (Devlin et al., 2018; Liu et al., 2020). Devlin et al. (2018) proposed BERT that leverages context from both left and right representations in each layer of the bidirectional transformer. The model is pretrained and released in the public domain, and can be trained in a simpler, yet efficient manner without having to make significant architectural changes for a specific task.

RoBERTa (Liu et al., 2020) is a replication study of BERT, trained on a dataset twelve times larger with bigger batches as compared to BERT. RoBERTa makes use of larger byte-pair encoding(BPE) vocabulary that helps to achieve better results than BERT on various downstream tasks. We use $RoBERTa_{large}$ and finetune it for the task using varying hyperparameters for different inputs.

We use the following three different types of input to study the effect of context on the performance of $Roberta_{large}$ to detect sarcasm:

1. **Response-only** : Input containing only the target response.
2. **Context-Response** : Input containing target response appended to the related context (containing previous responses).
3. **Context-Response (Separated)**: Input containing a separation token separating target response from context (containing previous responses).

Split/Dataset	Reddit	Twitter
Training	2.491	3.867
Testing	4.254	3.164

Table 1: The average number of contexts per post.

4 Experiments

4.1 Dataset

Two datasets containing an equal number of sarcastic and non-sarcastic responses were sampled from Twitter and Reddit by the authors of the shared task. The Twitter dataset comprises $5,000$ English tweets, and the Reddit dataset contains $4,400$ Reddit posts along with their context, which is an ordered list of dialogues. The sarcastic response, present in both datasets, is the reply to the last dialogue turn in the context list.

From Table 1, we infer that the dataset suffers from a significant mismatch in the average number of context responses between the training and testing split. This mismatch is particularly evident in the Reddit dataset, where the training split contains 1.71 times the number of contexts provided on average as compared to the testing split. We observe a similar mismatch, in the opposite direction, with the testing split containing 1.22 times the average number of contexts as compared to the training split. We argue that this mismatch in training and test splits in terms of context lengths, adds a layer of complexity to the problem. Consequently, we use the last two dialogues as the context in the *Context-Response* input and *the Context-Response (Separated)* input.

4.2 Experimental Setting

In this subsection, we outline the experimental setup for the sarcasm detection task and present the results obtained on the blind test set. For experiments, we used $RoBERTa_{large}$ having $355M$ parameters with a $50,265$ vocabulary size. For validation, we trained and evaluated our model for three

Input	F1-score	Precision	Recall
Response-only	0.752	0.752	0.753
Context-Response	0.772	0.772	0.772
Context-Response (Separated)	0.771	0.771	0.771

Table 2: Experimental Results for the Twitter test dataset.

Input	F1-score	Precision	Recall
Response-only	0.679	0.679	0.679
Context-Response	0.681	0.684	0.692
Context-Response (Separated)	0.716	0.716	0.718

Table 3: Experimental Results for the Reddit test dataset.

different inputs using a $90 - 10$ train-validation split.

We finetune $RoBERTa_{large}$ with a learning rate of $1 * 10^{-5}$ for 3 epochs. We used a sequence length of 50 for the *Response-only* input and 256 for the other two types of inputs, *Context-Response* input and *Context-Response (Separated)* input. We evaluate all our models on the following metrics: *F1 score*, *Precision-1* and *Recall-1*.

4.3 Results

In Table 2 and Table 3, we illustrate the effect of adding contextual information to the target response. We see an increase of 2.6% and 0.3% in *F1-score* upon including previous contexts with the response in the Twitter dataset and Reddit dataset respectively.

We also investigate the effect of adding a separation token between contextual information and the target response in the predictive performance of the model. We observe a 5.13% increase in *F1-score* in the Reddit dataset but a 0.1% decrease in *F1-score* in the Twitter dataset. We observe a similar pattern as the *F1-score* in both of its constituent metrics of *Precision* and *Recall*.

5 Conclusion

This work presents the importance of context while detecting sarcasm from responses sampled from Twitter and Reddit. Our proposed architecture using three different inputs performs competitively for both datasets showing that the addition of contextual information to target response improves the performance of the fine-tuned contextual word embeddings for detecting sarcasm. We further show that the addition of a separation token between context and target response also performs competitively, markedly showing an improvement of 5.13% in the *F1-score* of Reddit dataset. Future works include exploring different contextual cues, including user-specific attribute information and extending this hypothesis to other figurative speeches like irony detection and humor detection.

References

David Bamman and Noah A. Smith. 2015. Contextualized sarcasm detection on twitter. In *ICWSM*.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. pages 2506–2515.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Debanjan Ghosh, Alexander Fabbri, and Smaranda Muresan. 2018. Sarcasm analysis using conversation context. *Computational Linguistics*, 44:1–56.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon, USA. Association for Computational Linguistics.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Roberta: A robustly optimized bert pretraining approach. In *International Conference on Learning Representations*. Accepted.

Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marilyn Walker. 2017. Are you serious?: Rhetorical questions and sarcasm in social media dialog.

Kartikey Pant, Tanvi Dadu, and Radhika Mamidi. 2020. Towards detection of subjective bias using contextualized word embeddings. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 75–76, New York, NY, USA. Association for Computing Machinery.

Rossano Schifanella, Paloma Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms.

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *ACL*.

Orion Weller and Kevin D. Seppi. 2019. Humor detection: A transformer gets the last laugh. In *EMNLP/IJCNLP*.