When Do “More Contexts” Help with Sarcasm Recognition?

[†]Ojas Nimase and Sanghyun Hong
[†]Westview High School, Oregon State University
[†]ojasnimase@gmail.com, sanghyun.hong@oregonstate.edu

Abstract

Sarcasm recognition is challenging because it needs an understanding of the true intention, which is opposite to or different from the literal meaning of the words. Prior work has addressed this challenge by developing a series of methods that provide richer *contexts*, *e.g.*, sentiment or cultural nuances, to models. While shown to be effective individually, no study has systematically evaluated their collective effectiveness. As a result, it remains unclear to what extent additional contexts can improve sarcasm recognition. In this work, we explore the improvements that existing methods bring by incorporating more contexts into a model. To this end, we develop a framework where we can integrate multiple contextual cues and test different approaches. In evaluation with four approaches on three sarcasm recognition benchmarks, we achieve existing state-of-the-art performances and also demonstrate the benefits of sequentially adding more contexts. We also identify inherent drawbacks of using more contexts, highlighting that in the pursuit of even better results, the model may need to adopt societal biases.

1. Introduction

Sarcasm recognition carries importance in various domains, ranging from social media analysis (Amir et al., 2016) to product review classification (Parde and Nielsen, 2018). Beyond its practical applications, it also offers valuable insights into human behavior. For instance, Persicke et al. (2013) use sarcasm recognition to investigate the behaviors of individuals on the autism spectrum. But recognizing sarcasm is challenging because sarcastic expressions involve irony, are heavily context-dependent, and frequently depend on the tone of speeches (Parde and Nielsen, 2018).

Prior work addresses this challenge by integrating *more contexts*, typically sourcing additional information not readily discernible from the training corpus. Earlier work (Riloff et al., 2013) proposed learning representations (hereafter, we refer to as embeddings) that encode the positive or negative meaning of the words and use them to identify contrasts in a text. Recent work (Hazarika et al., 2018) focuses on encoding rich contextual information into sentence-level embeddings, e.g., by combining affective features (Babanejad et al., 2020) or by leveraging additional training corpus to have the embeddings learn contexts implicitly (Ahuja and Sharma, 2020; Liu et al., 2023a).

While these individual efforts have led to significant improvements in sarcasm recognition, there is a lack of a systematic study determining to what extent each approach is more effective. It thus remains unclear which methods one should prioritize in using. It is also unknown what the possibilities and impossibilities are: where the failures in sarcasm recognition are attributed and if we can address them by developing new methods.

Contributions. Our contributions are twofold:

First, we systematically analyze the effectiveness of providing additional contexts on embeddings in sarcasm recognition. To run this analysis, we design a framework to process additional contextual information existing work leverages when classifying sarcastic texts from non-sarcastic ones.

We apply four different approaches and evaluate their performances on three sarcasm recognition benchmarks: IAC-V1, IAC-V2, and Tweets (Oraby et al., 2016; Van Hee et al., 2018a). Our findings are: (1) by combining embeddings from the four methods, we achieve the state-of-the-art performance shown in the baselines. (2) sentence-level embeddings are more effective than word-level embeddings in sarcasm recognition. (3) when the embeddings are learned from datasets, potentially containing more sarcastic texts, they offer more improvements in recognition. (4) a training method, i.e., SimCLR (Chen et al., 2020), effective in learning better embeddings in other domains, offer negligible performance improvement.

Second, we conduct a manual analysis of the test-set samples, correctly classified (or incorrectly labeled) by each approach, and discuss the possibilities and impossibilities of sarcasm recognition. We observe that the samples are incorrectly classified initially become correctly classified after we provide more contexts. We also find the test-set samples where we fail to label correctly, even with all the embeddings combined. Surprisingly, from our manual analysis, we show that a model needs to learn societal biases to be correct in classifying these samples. Our result implies that models may need to learn undesirable biases or embeddings may require to encode them to further improve a model’s performance in sarcasm recognition.

LREC-COLING 2024, pages 17537–17543
20-25 May, 2024.

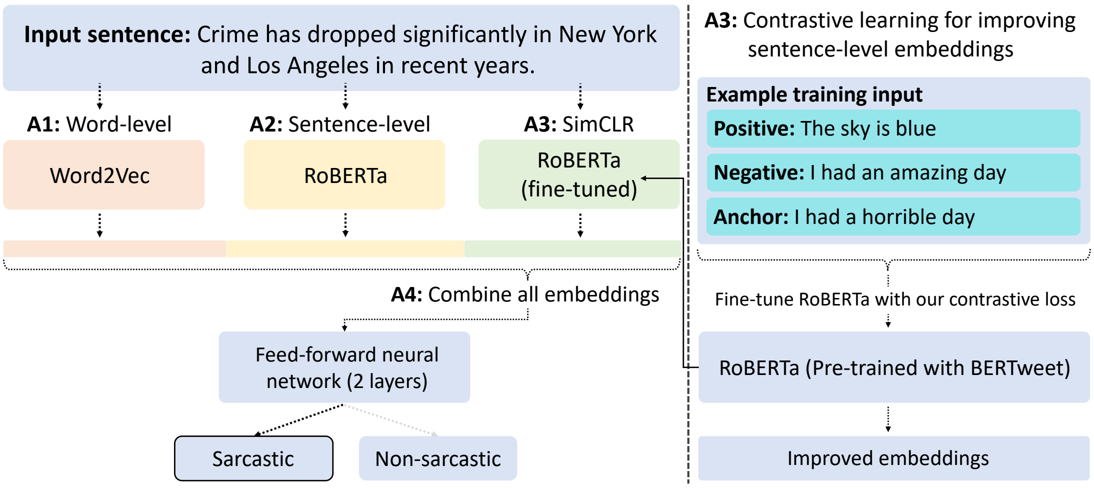
Input sentence: Crime has dropped significantly in New York and Los Angeles in recent years.
A1: Word-level
A2: Sentence-level
A3: SimCLR
Word2Vec
RoBERTa
RoBERTa (fine-tuned)
A4: Combine all embeddings
Feed-forward neural network (2 layers)
Sarcastic
Non-sarcastic
A3: Contrastive learning for improving sentence-level embeddings
Example training input
Positive: The sky is blue
Negative: I had an amazing day
Anchor: I had a horrible day
Fine-tune RoBERTa with our contrastive loss
RoBERTa (Pre-trained with BERTweet)
Improved embeddings

Figure 1: **Our framework.** We illustrate how the framework incorporates four different approaches and how we re-train sentence embeddings by adapting a contrastive learning technique (Chen et al., 2020).

2. A Framework for Our Study

Our goal is to study *possibilities and impossibilities* when we use more contexts in sarcasm recognition: how additional contexts have been improving the performance and what the limits are against pushing the state-of-the-arts. This section will introduce our method to answer these questions.

2.1. Methods That Offer More Contexts

We employ (and develop) four methods for incorporating additional contexts. The first two methods (A1 and A2) implement the representative approaches in prior work. The next one (A3) studied in different domains, but we adopted to sarcasm recognition. The last method (A4) is the combinations of A1–3, utilizing their embeddings at once.

A1: Word-level contexts. Initial work (Riloff et al., 2013) leverages word embeddings, such as Word2Vec or GloVe (Mikolov et al., 2013; Pennington et al., 2014) for identifying sarcasm. We implement this approach in our framework. Given a sarcastic text, we sum up the embeddings of the words in the text and feed them to a classifier to label if the text is sarcastic. The idea behind this approach is to quantify the contrast between the words. Positive words are likely to be near another positive word, and negative words do so; thus, the task becomes identifying if negative and positive words are combined together to deliver meanings different from literal meanings of words.

A2: Sentence-level contexts. The next component of our framework uses widely-used language models, based on transformer architectures, such as RoBERTa (Liu et al., 2019). These models generate sentence-level embeddings: they take a sentence (a sequence of words) and outputs a k-dimensional vector. The typical choice of k is 768. A standard practice of leveraging these models is to pre-train and fine-tune. We fine-tune a model, pre-trained on a large corpus of text data, such as BookCorpus (Zhu et al., 2015), on sarcasm recognition data. The intuition behind this is: even if the text, used to pre-train a model, is from domains different from sarcasm recognition, it may offer additional contexts to improve the performance.

A3: Improve sentence-level embeddings using contrastive learning. We train language models to learn sentence-level embeddings by maximizing (1) agreement between a non-sarcastic text and another unrelated non-sarcastic text, and (2) disagreement between the non-sarcastic text and its sarcastic translation. To learn such embeddings we adapt a popular contrastive learning framework (SimCLR), presented by Chen et al. (2020). Our loss function is formulated as follows:

$$\mathcal{L}_{i,j,k} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\exp(sim(z_i, z_j)/\tau) + \exp(sim(z_i, z_k)/\tau)}$$

where $sim(\cdot)$ is the cosine similarity, z_i, z_j, and z_k are the anchor, positive and negative embeddings, and τ is a temperature parameter. In our context, z_i is the non-sarcastic text, z_j is an unrelated non-sarcastic text, and z_k is a direct sarcastic translation text of the anchor non-sarcastic text. Re-training with the loss allows a model to encode the contexts that make non-sarcastic and sarcastic sentences different in the embedding space.

A4: Combine word- and sentence-level embeddings. We further combine the embeddings from the above approaches to leverage full contexts.

2.2. Putting All Together

We finally present a framework that enables us to individually (and also comprehensively) evaluate the effectiveness of our approaches. The architecture of our framework is shown in Figure 1.

Methods	IAC-V1				IAC-V2				Tweets			
	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
A1: Word2Vec	49.8	66.5	49.8	100.	54.9	68.5	52.7	98.1	39.7	56.8	39.7	100.
A2: RoBERTa	63.3	65.6	61.5	70.4	74.3	75.3	72.5	78.4	56.6	59.9	47.3	81.7
A2: BERTweet (RoBERTa)	54.9	37.4	60.6	27.0	75.0	75.1	74.4	74.8	63.8	64.0	52.8	81.0
A3: BERTweet (SimCLR)	58.3	47.0	64.1	37.1	75.2	75.2	75.2	75.1	62.8	63.4	52.0	81.4
A4: All Embeddings	72.1	72.1	71.9	72.3	**84.0**	83.2	**85.8**	80.8	**82.0**	**80.2**	71.3	**91.6**
Baselines (Liu et al., 2022)												
ADGCN-RoBERTa	**72.4**	**72.4**	72.5	**72.4**	82.1	82.1	82.2	82.1	72.2	71.4	71.3	71.9
DC-Net-RoBERTa	69.3	69.1	69.7	69.3	83.7	**83.7**	83.7	**83.7**	70.9	68.7	69.7	68.3
RoBERTa	72.1	71.9	**73.0**	72.1	82.7	82.7	82.9	82.9	72.7	72.8	**72.8**	73.9

Table 1: **Performance comparison** of four different approaches (A1–4) to encoding more contexts in sarcasm recognition. We compare accuracy, F1-score, precision, and recall. The bottom two rows are the performance from the baseline approach by Liu et al. (2022). Best results are highlighted in **bold**.

3. Evaluation

We now empirically evaluate the effectiveness of four different approaches (A1–4). We also analyze samples where each approach can improve upon.

3.1. Experimental Setup

Datasets. We run our experiments with three benchmarks: IAC-V1, IAC-V2, and Tweets (Oraby et al., 2016; Van Hee et al., 2018b), widely used in sarcasm recognition. Each sentence in the dataset is annotated as sarcasm or non-sarcasm, and we use them as labels. We additionally use SarcasmSIGN (Peled and Reichart, 2017), composed of sarcastic texts and their multiple, direct, non-sarcastic translations, for contrastive training of sentence embedding models. Note that SarcasmSIGN contains duplicates of non-sarcastic translations, and we filter them out before use.

Models. We harness the word embeddings produced by Word2Vec (Mikolov et al., 2013). To obtain sentence embeddings, we utilize pre-trained models available from Huggingface. Specifically, we use the RoBERTa-based models (Liu et al., 2019): roberta-base[1] and vinai/bertweet-base[2]. BERTweet models (Nguyen et al., 2020) undergo pre-training on English Tweets, enabling them to learn embeddings from more sarcastic texts.

Metrics. We measure the performance using the following four metrics: classification accuracy (or *accuracy*), *F1-score*, *precision*, and *recall*.

Our detailed experimental setup is in Appendix A.

3.2. Quantitative Evaluation

Table 1 summarizes our results. Overall, we find that using more contexts indeed helps with improving the performance in sarcasm recognition.

We first observe that, when all the embeddings are combined, we achieve the best performance. In IAC-V1/-V2, the performances are comparable, or it is better than the baselines (Liu et al., 2022) in Tweets. It is an interesting result because we achieve performances comparable to the baselines, designed explicitly for better sarcasm recognition, by simply combining more contexts. The results suggest that the improvements from the prior work may come from using more data, not from a delicate design of their methodology.

Now we turn our attention to how much performance improvement each approach brings. From the first row (A1), we see that word-level embeddings that encode the contexts from nearby words are not sufficient to perform well in sarcasm recognition. Sentence-level embeddings (A2) that capture contexts from long-range dependency significantly improve performance. The performance further increases when models are pre-trained on a corpus (A3), potentially including more sarcastic texts, such as English Tweets. Contrastive learning, in *contrast* to the advances made in other domains, does not improve the performance more. Instead, if we use all the embeddings, this straightforward approach leads us to the best (A4).

3.3. More Does Not Always Mean Better

We now manually analyze the samples correctly classified by an approach but misclassified by the preceding one. Previous studies have relied on amortized metrics, e.g., accuracy, to quantify performance improvements. While shown effective, they often leave ambiguity regarding whether the claimed improvements result from the proposed techniques or if other factors contribute to the increased accuracy. Our manual, per-sample anal-

[1] https://huggingface.co/roberta-base

[2] https://huggingface.co/vinai/bertweet-base

Methods	Example Texts	Pred.	Truth
A1: Word	I thought God forbid them to eat dead cows, or was it poultry? emoticonXRolleyes	S.	S.
	As a gun owner I'm also a property owner. Or are you denying that guns are property?	NS.	NS.
A2: RoBERTa	See, a terrorist attack is probably the sort of thing I would use as an excuse to not go to work...	S.	S.
	So why are you so afraid of it? If it is bad you will have a choice to go to a private insurance company.	NS.	NS.
A2: BERTweet (RoBERTa)	So everything that other people say on a website or in a book is just and opinion? And everything you say is a fact? Nice how you've got things set up.	S.	S.
	Please provide the actual estimates of the time required along with the necessary references if you please! How much time WOULD it take with confidence limits please!	NS.	NS.
A3: BERTweet (SimCLR)	This is just plain dumb. Abortion is NOT the primary means of birth control. If it is used as birth control, it's because others have failed or haven't been tried.	NS.	NS.
	It's a lot easier to kill someone with a gun than a cigarette or a beer.	S.	S.
A4: All Embeddings	Bravo, Penfold! You are the neatest pricker of balloons with the shortest of needles whom I have come across!	S.	S.
	The idea of abortion as population control is absurd, especially forced abortions as someone mentioned a few posts ago. Anyone who has read a biology book knows the world has methods of population control on its own, we don't need to be doing stuff like that ourselves.	NS.	NS.
A4: All Embeddings (Incorrect predictions)	The tactics pro-lifers use make the Nazis look like the little league. I mean, seriously. The reason we are dealing with terrorism is because women have the right to the abortion procedure. Wow. Please give me one way those two things relate to each other.	S.	NS.
	The VPC has a political agenda. The FBI? That is like saying I believe Coke taste better than Pepsi because the Coke commercial says so.	NS.	S.

Table 2: **Qualitative analysis.** Each successive method (except for the first and last methods) correctly classifies the samples incorrectly classified by the previous method, we underline the parts that seem to cause this performance improvement. S. signifies sarcastic text and NS. signifies non-sarcastic text.

ysis is a starting point to take an in-depth look at existing approaches and whether they are improving as shown in their original studies. We find in our qualitative analysis that in some cases, the improvements come as claimed, but in other cases, it is from undesirable model behaviors like biases. Examples are shown in Table 2.

A1 and A2. We show how sentence embeddings (A2) enhance a model's understanding of sarcastic texts. For example, the second two rows show that an approach solely relying on word embeddings cannot capture the long-range dependency between, e.g., "terrorist attack" and "not go to work." A1 classifies the example text as non-sarcasm. If sentence embeddings learn from texts potentially containing sarcasm (A2: BERTweet), the embeddings of "And everything you say is a fact?" are not similar to those of genuine questions. They are closer to accusatory questions.

A3. If the models that generate sentence embeddings are further fine-tuned using the contrastive learning approach, we observe that the embeddings begin to encode paraphrases commonly found in sarcastic texts, such as 'plain dumb.' But, in terms of performance, these advancements result in only a marginal difference (see Table 1 1).

A4. If we combine all the embeddings, we see the advantage of using information from both word-level and sentence-level embeddings. In word embeddings, the "neatest picker" and "shortest of needles" contain two words with an opposite sentiment. However, just looking at individual phrases is not sufficient to identify a sarcastic tone, and the sentence-level embeddings enable connecting the two and recognizing the sarcasm.

Biases. We further analyze the failure cases of A4 and find that, to make these sample texts correctly classified, embeddings may need to encode undesirable biases. For example, in the last two rows, to understand the sarcasm in the first text, a model may need to have negativity toward "pro-lifers" to align it with "Nazi." The same goes for the second example. A model (or embeddings) may need to be biased against conspiracy theorists to correctly classify the sample text as sarcasm.

More analyses can be found in Appendix B.

4. Conclusion

This paper studies the role of rich contextual information in sarcasm recognition. To conduct this study, we develop a framework that implements four representative approaches to incorporating richer contexts for sarcasm recognition. By evaluating these approaches on three sarcasm recognition benchmarks, we provide a new viewpoint on long-held beliefs in sarcasm detection. We show that: (1) Just combining more embeddings will offer the same performance in sarcasm detection as using complex model architectures or delicate training methods. (2) Pushing the performance further may require a model to learn undesirable biases, necessitating rethinking whether we should keep improving the current approaches.

What's Next? Our work underscores the need for future work to develop new methodologies for building models that excel in sarcasm detection and minimize reliance on undesirable biases, such as those related to gender or societal norms. To achieve this goal, we encourage the following directions for future research: (1) A systematic investigation of when and how these biases are introduced to a model. This involves adapting existing metrics or devising new ones to accurately quantify biases present in models. Moreover, we envision developing a novel method to determine which training instances significantly impact the accurate identification of specific sarcastic expressions. (2) In light of the remarkable capabilities of large-language models, future research should assess whether increasing model size effectively addresses the bias issues we have identified. Although our manual analysis suggests that improvements might inadvertently depend on undesirable biases, the efficacy of scaling as a solution remains uncertain. It is therefore important to empirically test this hypothesis to determine the viability of scaling as a strategy to mitigate bias. (3) Moreover, future work may focus on the cross-collaboration that must occur between research and social institutions. Given that enhanced sarcasm detection may inadvertently learn and propagate undesirable biases, it is important to prevent the deployment of such biased models within social institutions. Moreover, considering that much of the training data for sarcasm detection models comes from social media (e.g., Twitter), there is a need for researchers to collaborate with these companies to limit the introduction of harmful data. By working together, we hope to develop strategies that ensure the ethical and unbiased development of natural language processing methods.

5. Acknowledgements

We thank anonymous reviewers for their valuable feedback. This work is partially supported by the Samsung Global Research Outreach (GRO) Program and the Google Faculty Research Award.

6. Bibliographical References

Ravinder Ahuja and S. C. Sharma. 2020. A review paper on sarcasm detection. In *International Conference on Artificial Intelligence: Advances and Applications 2019*, pages 371–381, Singapore. Springer Singapore.

Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mário J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 167–177, Berlin, Germany. Association for Computational Linguistics.

Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 225–243, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

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.

Yiyi Liu, Yequan Wang, Aixin Sun, Xuying Meng, Jing Li, and Jiafeng Guo. 2022. A dual-channel framework for sarcasm recognition by detecting sentiment conflict. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1670–1680, Seattle, United States. Association for Computational Linguistics.

Yiyi Liu, Ruqing Zhang, Yixing Fan, Jiafeng Guo, and Xueqi Cheng. 2023a. Prompt tuning with contradictory intentions for sarcasm recognition. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 328–339, Dubrovnik, Croatia. Association for Computational Linguistics.

Yiyi Liu, Ruqing Zhang, Yixing Fan, Jiafeng Guo, and Xueqi Cheng. 2023b. Prompt tuning with contradictory intentions for sarcasm recognition. In *European Chapter of the ACL*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

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.

Natalie Parde and Rodney Nielsen. 2018. Detecting sarcasm is extremely easy ;-). In *Proceedings of the Workshop on Computational Semantics beyond Events and Roles*, pages 21–26, New Orleans, Louisiana. Association for Computational Linguistics.

Lotem Peled and Roi Reichart. 2017. Sarcasm SIGN: Interpreting sarcasm with sentiment based monolingual machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1690–1700, Vancouver, Canada. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Angela Persicke, Jonathan Tarbox, Jennifer Ranick, and Megan St. Clair. 2013. Teaching children with autism to detect and respond to sarcasm. *Research in Autism Spectrum Disorders*, 7(1):193–198.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

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, Seattle, Washington, USA. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018a. 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.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018b. We usually don't like going to the dentist: Using common sense to detect irony on Twitter. *Computational Linguistics*, 44(4):793–832.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

A. Experimental Setup in Detail

Here we describe our experimental setup in detail for the reproducibility of our analysis results. Our code is available at https://github.com/secure-ai-systems-group/sarcasm-detection.

A1: Word-level embeddings. We employ the Word2Vec model from Gensim (Řehůřek and Sojka, 2010) to generate text embeddings. We first convert each word within a text into its corresponding embeddings and concatenate them. We limit each text sample to a maximum of 50 words.

A2: Sentence-level embeddings are generated by using the `vinai/bertweet-base` and `roberta-base` models. We use the AutoTokenizer class for our models. We feed the token-level embeddings, generated by the tokenizer to our models. We average the embeddings obtained from our model to make a single sentence embedding.

A3: Contrastive-learning process. To further improve the quality of sentence-level embeddings, we fine-tune the vinai/bertweet-base model via our contrastive learning technique:

(1) We first fine-tune the `vinai/bertweet-base` model on the SarcasmSIGN dataset (Peled and Reichart, 2017). Before fine-tuning, we remove duplicates from the dataset.
(2) We adapt the contrastive learning framework presented by Chen et al. (2020) for visual representations to enhance our sentence-level embeddings. The framework uses a 2-layer feedforward neural network to produce 256-dimensional representations; we follow this process and decrease the representation dimension from 768 to 256. We then fine-tune these two models using *NT-Xent* loss.
(3) We fine-tune the model for 10 epochs, using a temperature of 0.7, batch size of 50, a learning rate of 1e-5, a weight decay of 1e-3, and the AdamW optimizer.
(4) We follow the same process outlined in A2 with this fine-tuned `vinai/bertweet-base` to create the sentence-level embeddings.

Sarcasm recognition models. We employ a 2-layer feedforward neural network to implement our models. When we test each approach individually, we set the dimension of the input linear layer to 768. We set the hidden layer dimension to 128 and employ the ReLU activation function for non-linearity. The final linear layer classifies the text as sarcastic or non-sarcastic. We train them for 5 epochs using a weight decay of 0.01, a learning rate of 1e-5, and a batch size of 32. We use the cross-entropy loss and AdamW optimizer.

When all four types of embeddings are used, we concatenate them, and therefore, the input layer's dimension increases from 768 to 39936. The A1, A2 RoBERTa, and A3 embeddings have already been generated and are fed in as lists while A2 BERtweet embeddings are added by incorporating 'vinai/bertweet-base' with the model. A linear layer is used to reduce this dimensionality back to 768 and the other architectural choices are kept the same. We train this model for 5 epochs with a batch size of 16, a weight decay of 0.01, and a learning rate of 1e-5. We employ the F-β score as our loss function. We use the AdamW optimizer.

B. More Qualitative Analysis

Here we provide more examples of a model learning biases for improving sarcasm recognition.

(1) Katie pisses me off so bad #TheApprentice
(2) @cnsnews Obama and Hillary convinced Ukraine that they would protect them if they essentially disarm. Need to keep at least one promise.
(3) Everytime I try to like Chris Brown he does something to royally eff that up. Dude is a chronic loose cannon #chrisbrown #Karrueche
(4) Again, as an ignorant layman, I can only get the gist of this material, but how anyone could possibly argue against the genetic code as a product of intelligent design is beyond me.

Examples correctly classified by a model potentially leaned societal biases. We showcase four examples, incorrectly classified by the models in A1–3, while correctly classified by our A4 model.

We showcase example texts (1)–(3) that were incorrectly classified as sarcastic by A1–A3 and correctly classified by A4 as non-sarcastic. (4) was incorrectly classified as non-sarcastic by A1-A3 and correctly classified as sarcastic by A4. We conduct a manual analysis on why: (1) shows A4 may become biased against Katie, a contestant in "#TheApprentice". (2) shows A4 may become biased against "Obama and Hilary" to correctly classify it. (3) shows A4 may become biased against "Chris Brown". (4) shows A4 may become biased against "Intelligent Design" and mock it.