SimSCLSD: A Simple Framework for Supervised Contrastive Learning of Sarcasm Detection

반어 표현 탐지를 위한 단순 지도 대조 학습 프레임워크

2026년 2월

서울대학교 공과대학

전기 · 정보공학부

엄 태 윤

Abstract

Sarcasm detection constitutes a pivotal challenge in Natural Language Processing (NLP) due to its ubiquity in online discourse and its tendency to invert the polarity of literal statements. Failing to detect sarcastic intent often results in significant errors in downstream tasks such as sentiment analysis and opinion mining. However, sarcasm is highly nuanced and context-dependent, making it difficult for standard classification models to identify accurately. While pre-trained Transformer-based language models have established standard baselines, they are typically fine-tuned using a Cross-Entropy (CE) loss. This approach may not be optimal for learning discriminative representations, especially for a nuanced task like sarcasm where the boundary between classes is subtle.

To address this limitation, we present SimSCLSD, a Simple framework for Supervised Contrastive Learning of Sarcasm Detection, which introduces a two-stage training process. Contrastive learning is a paradigm that learns representations by contrasting positive pairs against negative pairs to induce a well-structured embedding space. In our first stage, we employ a supervised contrastive pre-training phase where a RoBERTa-based encoder is optimized to pull representations of same-label examples together while pushing apart representations of different-label examples. Unlike standard self-supervised contrastive methods that rely on data augmentation of a single instance, our approach leverages label information to utilize multiple positive samples within a batch. In the second stage, the model is fine-tuned with a standard CE loss to learn the final classification boundary.

We evaluate our approach on the Reddit and Twitter datasets from the FigLang 2020 Sarcasm Detection shared task. Our experiments demonstrate that this contrastive pre-training step can effectively create more separable representations, showing its potential as an effective intermediate step for fine-tuning transformers on nuanced, context-dependent classification tasks.

keywords: sarcasm detection, supervised contrastive learning, representation learning, text classification, conversational context, embedding space

Student Number: 2019-18535

Contents

1 Introduction

With the proliferation of social media platforms such as Twitter and Reddit, the volume of user-generated content has grown exponentially. This data offers immense potential for analyzing public opinion and sentiment. However, the informal nature of online discourse introduces significant noise, with figurative language—specifically sarcasm—posing a formidable barrier to accurate automated analysis [1]. Sarcasm is defined as a form of verbal irony where the speaker's intended meaning is often the opposite of the literal interpretation of their words. For instance, the phrase "Great weather we're having" spoken during a storm conveys a negative sentiment despite the positive adjective "Great." Consequently, failure to detect sarcasm can result in the complete inversion of a sentiment analysis system's prediction, rendering it unreliable for real-world applications.

The primary difficulty in sarcasm detection lies in its reliance on context. An utterance that appears positive in isolation may be revealed as sarcastic only when juxtaposed with prior conversational turns or specific user attributes. Recent advancements in Natural Language Processing (NLP) have seen the widespread adoption of Pre-trained Language Models such as BERT [2] and RoBERTa [3]. These models have achieved state-of-the-art performance on sarcasm detection benchmarks, most notably the FigLang 2020 Shared Task [1], where Transformer-based systems dominated the leaderboard by leveraging conversational history. These models typically employ a fine-tuning paradigm where the model is optimized using the Cross-Entropy (CE) loss function.

Despite their success, standard fine-tuning with CE loss has inherent limitations. CE loss focuses on maximizing the likelihood of the correct class but does not explicitly incentivize the model to learn high-quality, discriminative representations where samples of the same class form tight clusters [4]. In tasks with subtle decision boundaries, such as sarcasm detection, this can lead to representations that are poorly separated, reducing the model's generalization capability and robustness to noise.

To overcome these limitations, this study proposes SimSCLSD, a Simple framework for Supervised Contrastive Learning of Sarcasm Detection. Inspired by recent successes in computer vision [4] and sentiment analysis [5], we hypothesize that disentangling the representation learning phase from the classification phase can yield superior performance. Our approach utilizes Supervised Contrastive Learning (SupCon) as an intermediate training step. By explicitly optimizing the encoder to minimize intra-class variance and maximize inter-class variance of the feature representations, we generate an embedding space that is structurally more conducive to linear classification.

Our specific contributions are as follows:

1. We propose a two-stage training framework that integrates supervised contrastive learning with standard fine-tuning for the task of context-aware sarcasm detection.

2. We adapt the generic supervised contrastive loss for NLP by utilizing same-label sampling rather than data augmentation as the primary source of positive pairs, addressing the discrete nature of text data.

3. We conduct extensive experiments on the FigLang 2020 Reddit and Twitter datasets, demonstrating significant performance improvements over standard RoBERTa baselines.

2 Related Work

Our research builds upon three foundational pillars: the role of context in sarcasm detection, the application of Transformer architectures to this domain, and the emerging paradigm of supervised contrastive learning in NLP.

2.1 Sarcasm Detection and Contextual Dependency

Early approaches to sarcasm detection treated the problem primarily as a single-sentence classification task, relying on lexical cues and pattern matching [1, 6]. However, subsequent research has established that conversational context is indispensable. Wallace et al. (2014) [7] demonstrated that humans often require context to identify ironic intent, implying that computational models share this requirement. Recognizing this, the FigLang 2020 Shared Task on Sarcasm Detection was organized to benchmark systems on their ability to leverage conversational history from platforms like Reddit and Twitter [1]. The shared task highlighted that models incorporating preceding dialogue turns consistently outperformed those analyzing responses in isolation.

2.2 Transformers in Sarcasm Detection

Following the FigLang 2020 shared task, Transformer-based models became the de facto standard for sarcasm detection. Participants overwhelmingly adopted BERT and RoBERTa architectures, focusing on various methods to integrate context [1].

A prevalent strategy is input concatenation. Dong et al. (2020) [6] demonstrated that concatenating the context and target response into a single sequence significantly outperforms target-oriented models. Pant and Dadu (2020) [8] further refined this by showing that inserting explicit separation tokens between the context and response yields performance gains, particularly on the Reddit dataset.

Other researchers focused on the optimal length of context. Jaiswal (2020) [9] and Lee et al. (2020) [10] observed that utilizing the most recent three turns of dialogue provided the best balance between context and noise, with Lee et al. employing a context ensemble strategy to combine predictions from models trained on varying context lengths. Additionally, Ataei et al. (2020) [11] adapted Aspect-Based Sentiment Analysis (ABSA) architectures, treating the context as the "aspect" to attend to. While these works focused on architectural modifications and input formatting, our work diverges by optimizing the underlying loss function used for representation learning.

2.3 Supervised Contrastive Learning

The Cross-Entropy (CE) loss, while ubiquitous, is known to suffer from poor margins and a lack of robustness to noisy labels. Contrastive Learning has emerged as a robust alternative, popularized by self-supervised frameworks like SimCLR [12] in computer vision. The core objective is to map positive pairs (e.g., augmented views of an image) to nearby points in embedding space while pushing negative pairs apart. Khosla et al. (2020) [4] extended this to the fully supervised setting (SupCon), leveraging label information to include all samples of the same class as positive pairs. This formulation forces the model to learn tighter class clusters than CE loss alone.

Adapting this paradigm to NLP, Moukafih et al. (2022) introduced SimSCL, a simple supervised contrastive learning framework for text classification [5]. Unlike self-supervised approaches that rely on data augmentation to generate positive pairs, SimSCL posits that sentences belonging to the same class are positive examples of each other. They proposed a novel fully-supervised contrastive loss function designed to maximize inter-class distances while minimizing intra-class variance. Formally, the loss $\mathcal{L}_{\text{SupCon}}$ is defined as:

$$\mathcal{L}_{\text{SupCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}$$

where z_i is the normalized embedding of the anchor sample i, $P(i)$ is the set of indices for samples belonging to the same class as i (positives), $N(i)$ is the set of indices for samples belonging to different classes (negatives), $A(i)$ is the set of all other samples, and τ is the temperature parameter. This loss explicitly encourages the model to pull all samples of the same semantic class together in the embedding space, creating a more discriminative structure than standard cross-entropy training [5]. Our work applies this objective to the nuanced task of sarcasm detection, investigating its ability to disentangle complex semantic incongruities.

3 Proposed Method

Our approach, SimSCLSD, is a two-stage training framework designed to enhance context-aware sarcasm detection. The central premise is to decouple representation learning from classification.

3.1 Model Architecture

The model architecture is trained using a sequential two-stage process, which is illustrated in Figures 1 and 2. This separation allows the model to first learn the semantic structure of sarcasm before attempting to define a decision boundary.

3.1.1 Stage 1: Supervised Contrastive Pre-training

As illustrated in Figure 1, the first stage is dedicated to representation learning. The process begins with a batch of labeled inputs formatted as `context [SEP] response` [8]. These inputs are fed into the RoBERTa encoder. Unlike standard classification tasks that utilize the `[CLS]` token, this stage utilizes a mean-pooling operation over the token embeddings. This is because the `[CLS]` token is often biased towards next-sentence prediction tasks from pre-training, whereas mean-pooling aggregates information from the entire sequence, producing a more robust holistic representation for clustering objectives [13].

These representations are projected into a normalized embedding space. As depicted in the "Embedding Space" block of Figure 1, the Supervised Contrastive (SupCon) loss explicitly restructures this space [5]. The red circles represent sarcastic embeddings (h_{sarc}) and the blue squares represent non-sarcastic embeddings (h_{norm}). The objective function applies attractive forces to pull samples of the same class together (e.g., h_{sarc} to h_{sarc}) and repulsive forces to push different classes apart (e.g., h_{sarc} from h_{norm}). The dashed red arrow indicates the backpropagation path, where the encoder weights are updated to minimize this contrastive loss, resulting in a more discriminative feature space [4].

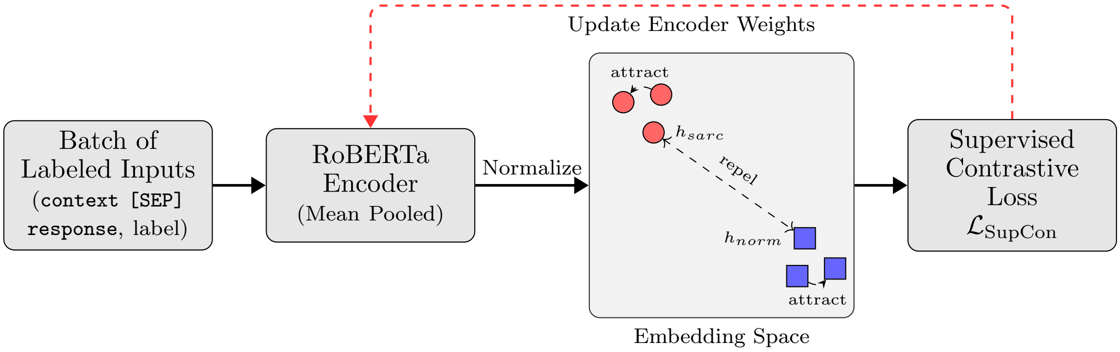
Update Encoder Weights
Batch of Labeled Inputs (context [SEP] response, label)
RoBERTa Encoder (Mean Pooled)
Normalize
attract
h_{sarc}
repel
h_{norm}
attract
Embedding Space
Supervised Contrastive Loss $\mathcal{L}_{\text{SupCon}}$

Figure 1: Overview of Stage 1: Supervised Contrastive Pre-training.

3.1.2 Stage 2: Classification Fine-tuning

Figure 2 details the second stage, where the model is fine-tuned for the binary classification task. The RoBERTa encoder is initialized with the weights optimized in Stage 1. In this stage, we switch to using the `[CLS]` token embedding, which is fed into a newly initialized linear classification head [4].

The model outputs a prediction ("Pred") which is compared against the "True Label" using a standard Cross-Entropy Loss. A critical component of this stage, visualized by the split update arrows in Figure 2, is the use of differential learning rates. The red arrows indicate that the classifier head is updated with a higher learning rate to rapidly learn the decision boundary, while the encoder is fine-tuned with a significantly lower learning rate. This strategy preserves the cluster structure learned in Stage 1 while allowing the model to adapt to the specific classification objective [5].

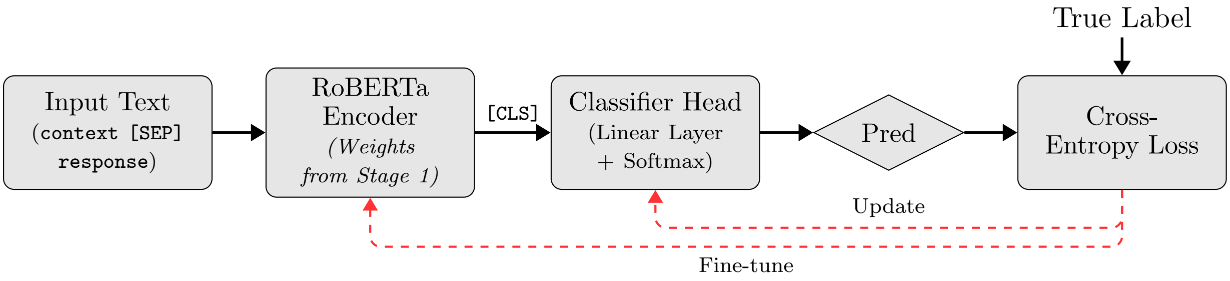
True Label
Input Text (context [SEP] response)
RoBERTa Encoder (Weights from Stage 1)
[CLS]
Classifier Head (Linear Layer + Softmax)
Pred
Cross-Entropy Loss
Update
Fine-tune

Figure 2: Overview of Stage 2: Classification Fine-tuning.

3.2 Input Formatting

Adopting the findings of Pant and Dadu (2020) [8], we format the input by concatenating the conversational context sequence and the target response, separated by special tokens. The input sequence is constructed as follows:

```
[CLS] context_1 [SEP] ... [SEP] target_response [EOS]
```

This format is effective since the Transformer architecture utilizes a self-attention mechanism. By placing both context and response in the same sequence, self-attention allows every token in the response to directly attend to every token in the context layers. This enables the model to effectively capture the context incongruity which is the hallmark of sarcasm.

4 Experiments

4.1 Datasets

We conducted our evaluation using the two primary datasets from the FigLang 2020 Shared Task on Sarcasm Detection [1]. Both datasets are balanced binary classification tasks and include conversational context. The statistics of context length for both datasets are visualized in Figure 3.

- **FigLang-Reddit:** Comprises 4,400 training samples and 1,800 test samples. It is derived from the self-annotated Reddit corpus by Khodak et al. (2018) [14], where users mark sarcasm using the `/s` tag. As shown in Figure 3, the dataset is characterized by short context lengths, with over 60% of samples containing only 2 prior turns.

- **FigLang-Twitter:** Comprises 5,000 training samples and 1,800 test samples. Sarcastic tweets are collected using hashtags such as #sarcasm and #sarcastic. Unlike Reddit, the non-sarcastic examples are not random tweets but are sentiment-bearing tweets (containing hashtags like #happy, #sad), making the discrimination task considerably harder [1]. Figure 3 illustrates that Twitter threads are longer, with a notable distribution of samples having more than 5 context turns.

For our experiments, we partitioned the provided training data into a 90% training set and a 10% validation set to monitor convergence and perform model selection.

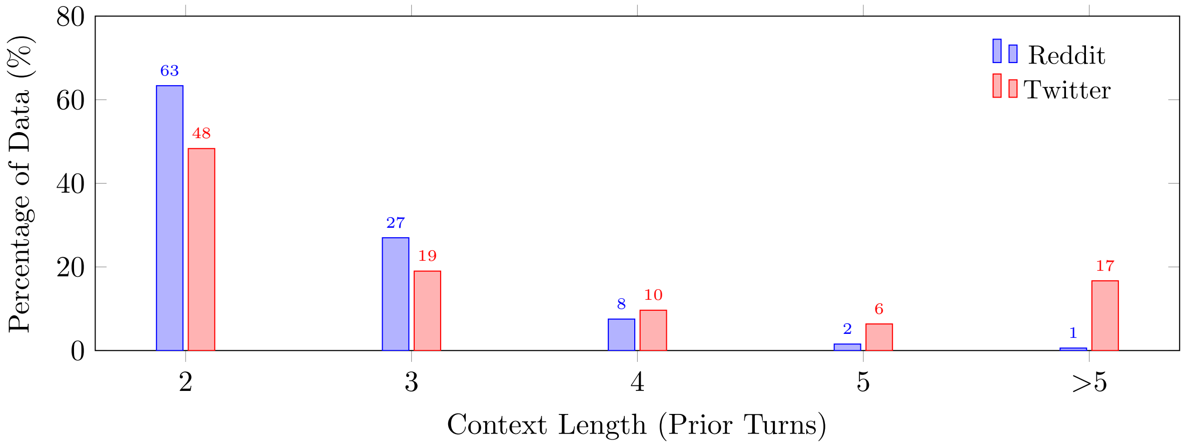

Figure 3: Distribution of context lengths in the training datasets.

4.2 Experimental Setup

We benchmarked our SimSCLSD framework against a strong baseline to isolate the effect of contrastive pre-training. All experiments utilized the `roberta-base` [3] architecture with a batch size of 64. Early stopping was applied with a patience of 2 epochs based on validation Macro F1-score.

Baseline (Standard Fine-Tuning): The model is fine-tuned directly on the target datasets using Cross-Entropy loss for up to 10 epochs. This represents the standard industry approach to text classification. Learning rates were tuned specifically for each dataset (3×10^{-5} for the classifier head; 5×10^{-7} to 1×10^{-6} for the encoder).

SimSCLSD (Ours): During the first stage, the encoder is trained for 20 epochs using $\mathcal{L}_{\text{SupCon}}$ with a learning rate of 5×10^{-5} and temperature $\tau = 0.2$. In the second stage, The model is fine-tuned using CE loss for up to 10 epochs. We used differential learning rates: 3×10^{-5} (Twitter) and 1×10^{-5} (Reddit) for the classifier, and 5×10^{-7} (Twitter) and 1×10^{-6} (Reddit) for the encoder. Dropout was set to 0.1 for Twitter and 0.5 for Reddit to mitigate overfitting given the dataset sizes.

4.3 Results and Discussion

Table 1 presents the performance metrics (Precision, Recall, and Macro F1) on the official test sets. The model checkpoint that performed best on the validation set was used for final testing.

Table 1: Performance comparison between the Baseline and the SimSCLSD framework on FigLang 2020 datasets. All metrics are Macro-averaged.

Dataset	**Split**	**Model**	**Precision**	**Recall**	**Macro F1**
Reddit	Dev	Baseline	0.5679	0.5568	0.5379
		SimSCLSD	**0.6974**	**0.6955**	**0.6947**
Reddit	Test	Baseline	0.5244	0.5228	0.5148
		SimSCLSD	**0.6166**	**0.6139**	**0.6116**
Twitter	Dev	Baseline	0.7352	0.7300	0.7285
		SimSCLSD	**0.8486**	**0.8480**	**0.8479**
Twitter	Test	Baseline	0.7047	0.6861	0.6788
		SimSCLSD	**0.7477**	**0.7461**	**0.7457**

On the FigLang-Reddit dataset, SimSCLSD achieves a Macro F1-score of 0.6116, outperforming the baseline (0.5148) by a substantial margin of 9.68 percentage points. Similarly, on the FigLang-Twitter dataset, our method yields an F1-score of 0.7457 compared to the baseline's 0.6788, an improvement of 6.69 percentage points.

These gains can be attributed to the nature of the contrastive loss. In sarcasm detection, the distinction between a sarcastic and a sincere statement often hinges on subtle semantic cues within the context. Standard CE training may struggle to separate these embeddings when the vocabulary overlap is high. By enforcing a compact clustering of sarcastic examples during pre-training, SimSCLSD essentially teaches the encoder to recognize the latent structure of sarcasm before it attempts to classify it, leading to a more robust decision boundary in the second stage.

5 Conclusion

In this paper, we introduced SimSCLSD, a two-stage framework for context-aware sarcasm detection that integrates supervised contrastive learning with transformer-based fine-tuning. We hypothesized that the widely used Cross-Entropy loss is suboptimal for nuanced linguistic tasks and that explicitly structuring the embedding space could yield better performance.

Our extensive experiments on the FigLang 2020 shared task datasets validated this hypothesis. By preceding standard classification training with a supervised contrastive pre-training phase, we achieved statistically significant F1-scores improvements of 9.7%p on the Reddit dataset and 6.7%p on the Twitter dataset. This confirms that learning to discriminate between classes in the representation space provides a stronger foundation for the final classifier than learning the decision boundary alone.

Future work could explore the applicability of this framework to larger architectures and investigate its generalization to other complex figurative language tasks, such as irony detection and metaphor identification.

References

[1] D. Ghosh, A. Vajpayee, and S. Muresan, "A report on the 2020 sarcasm detection shared task," *arXiv preprint arXiv:2005.05814*, 2020.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[4] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[5] Y. Moukafih, A. Ghanem, K. Abidi, N. Sbihi, M. Ghogho, and K. Smaïli, "Simscl: A simple fully-supervised contrastive learning framework for text representation," in *Australasian Joint Conference on Artificial Intelligence.* Springer, 2022, pp. 728–738.

[6] X. Dong, C. Li, and J. D. Choi, "Transformer-based context-aware sarcasm detection in conversation threads from social media," *arXiv preprint arXiv:2005.11424*, 2020.

[7] B. C. Wallace, L. Kertz, E. Charniak *et al.*, "Humans require context to infer ironic intent (so computers probably do, too)," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 512–516.

[8] K. Pant and T. Dadu, “Sarcasm detection using context separators in online discourse,” *arXiv preprint arXiv:2006.00850*, 2020.

[9] N. Jaiswal, “Neural sarcasm detection using conversation context,” in *Proceedings of the second workshop on figurative language processing*, 2020, pp. 77–82.

[10] H. Lee, Y. Yu, and G. Kim, “Augmenting data for sarcasm detection with unlabeled conversation context,” *arXiv preprint arXiv:2006.06259*, 2020.

[11] T. Shangipour ataei, S. Javdan, B. Minaei-Bidgoli *et al.*, “Applying transformers and aspect-based sentiment analysis approaches on sarcasm detection,” in *Proceedings of the second workshop on figurative language processing*, 2020, pp. 67–71.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[13] B. Li, F. Xia, Y. Weng, X. Huang, and B. Sun, “Simclad: A simple framework for contrastive learning of acronym disambiguation,” *arXiv preprint arXiv:2111.14306*, 2021.

[14] M. Khodak, N. Saunshi, and K. Vodrahalli, “A large self-annotated corpus for sarcasm,” in *proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.