A Match Made in Heaven: A Multi-task Framework for Hyperbole and Metaphor Detection

Naveen Badathala*, **Abisek Rajakumar Kalarani***, **Tejpalsingh Siledar***,
Pushpak Bhattacharyya
Department of Computer Science and Engineering, IIT Bombay, India
{naveenbadathala, abisekrk, tejpalsingh, pb}@cse.iitb.ac.in

Abstract

Hyperbole and metaphor are common in day-to-day communication (*e.g.*, "I am in deep trouble": how does *trouble* have *depth*?), which makes their detection important, especially in a conversational AI setting. Existing approaches to automatically detect metaphor and hyperbole have studied these language phenomena independently, but their relationship has hardly, if ever, been explored computationally. In this paper, we propose a multi-task deep learning framework to detect hyperbole and metaphor simultaneously. We hypothesize that metaphors help in hyperbole detection, and vice-versa. To test this hypothesis, we annotate two hyperbole datasets- *HYPO* and *HYPO-L*- with metaphor labels. Simultaneously, we annotate two metaphor datasets- *TroFi* and *LCC*- with hyperbole labels. Experiments using these datasets give an improvement of the state of the art of hyperbole detection by $\sim \mathbf{12\%}$. Additionally, our multi-task learning (MTL) approach shows an improvement of up to $\sim \mathbf{17\%}$ over single-task learning (STL) for both hyperbole and metaphor detection, supporting our hypothesis. To the best of our knowledge, ours is the first demonstration of computational leveraging of linguistic intimacy between metaphor and hyperbole, leading to showing the superiority of MTL over STL for hyperbole and metaphor detection[1].

1 Introduction

The use of figurative language is very common in natural discourse, and it is reflected in the content generated in social media networks (Abulaish et al., 2020). Figurative languages are used to establish some communicative goals such as expressing a negative emotion, drawing attention to a part of the text, or adding interest to a subject. (Roberts and Kreuz, 1994). The understanding of figurative languages like sarcasm, metaphor, simile, irony, and hyperbole is crucial for many NLP tasks such as building accurate sentiment analysis systems or developing conversational AI systems that can hold meaningful conversations (Figure 1). This has led to great interest and value in understanding these figurative languages. Figurative languages like metaphor (Rai and Chakraverty, 2020) and sarcasm (Joshi et al., 2017) are studied extensively while hyperbole remains less explored.

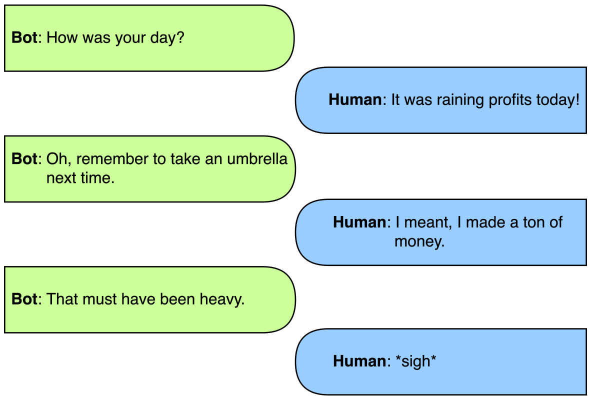

Figure 1: An example of the need for detecting hyperbolic and metaphoric sentences for AI systems.

Metaphor is the most common choice of figurative language, while hyperbole is the second most adopted rhetorical device in communication (Roger J., 1996) and hence it is important to study and process them automatically. Hyperbole is a figurative language that uses exaggeration to emphasize a point, while metaphor makes a comparison between two things to indicate a resemblance.

1.1 Motivation

Relevance theorists had long treated both metaphors and hyperboles as not genuinely distinct categories as they are very closely related to each other (Sperber and Wilson, 2008). Recent research has highlighted the distinctive features of hyperboles over metaphors (Carston and Wearing,

*Equal contribution.
[1]Code and data are available at: https://github.com/abisekrk/multitask_hyperbole_metaphor_detection

Findings of the Association for Computational Linguistics: ACL 2023, pages 388–401
July 9-14, 2023

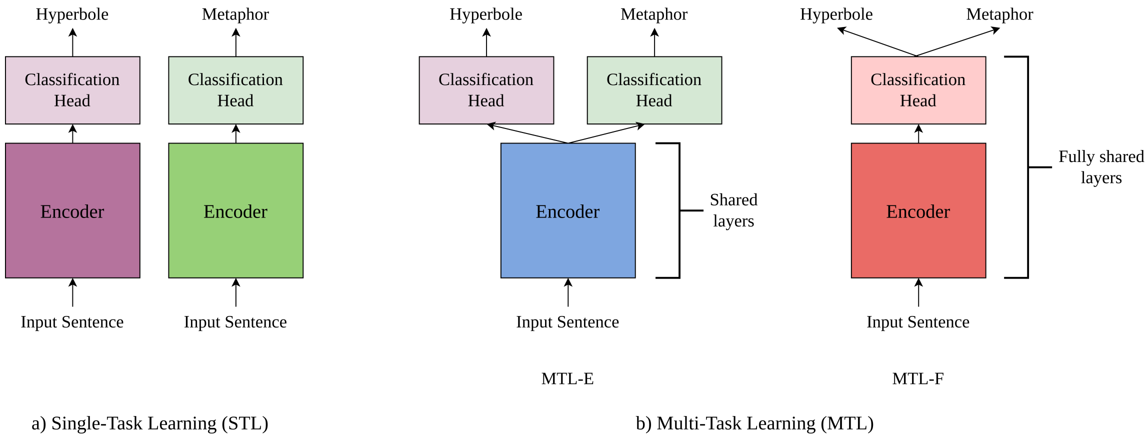

Figure 2: Overview of the single-task and multi-task learning architectures for detecting hyperbole and metaphor. a) Single-Task Learning (STL) model. b) Two variants of the Multi-Task learning (MTL) model: Multi-task learning with shared encoders (MTL-E) model and Multi-Task Learning with fully shared layers (MTL-F) model.

2015). However, on the computational side, hyperbole and metaphor detection have been treated as isolated problems so far.

Both metaphors and hyperboles use figurative elements to express an idea rather than presenting them literally, but this linguistic insight hasn't been exploited computationally in previous works. We hypothesize that this shared characteristic can be captured at the embedding level by training transformer models to learn these representations jointly using multi-task learning. Existing metaphor detection systems focus on identifying metaphoricity at the token-level, whereas hyperbole detection systems focus on sentence-level classification. In our work, we highlight the effectiveness of performing sentence-level classification for both hyperboles and metaphors in a multi-task setting.

1.2 Contributions

Our contributions are:

1. Extensions to the existing datasets amounting to **16, 024** sentences which include,
   (a) HYPO and HYPO-L datasets annotated with metaphor labels.
   (b) TroFi and LCC datasets annotated with hyperbole labels.
2. Demonstration of the superiority of multi-tasking over single-tasking for hyperbole and metaphor detection.
3. State-of-the-art results for sentence-level hyperbole detection on the HYPO dataset (F1 score- **0.881**).

Sentence	**Hyperbole**	**Metaphor**
Your plan is too risky, its a suicide	✓	✓
This kind of anger rages like a sea in a storm	✓	✗
Her strength awoke in poets an abiding love	✗	✓
My ex boyfriend! Treacherous person	✗	✗

Figure 3: Example sentences with Hyperbole and Metaphor labels.

4. Benchmark results for sentence-level metaphor detection on our label-balanced LCC dataset (F1 score- **0.805**).

2 Background and Definitions

Metaphor Metaphor is a literary device that uses an implicit comparison to drive home a new meaning. Metaphors consist of a source and target domain in which the features from the source domain are related to the features in the target domain through comparable properties (Lakoff, 1993). For instance, *"Life is a journey,"* implies a comparison between life and journey through the idea of having a beginning and an end. In this work, we do not consider similes as metaphors as they make an explicit comparison.

Hyperbole Hyperbole is a figurative language in which the literal meaning is exaggerated intentionally. It exaggerates expressions and blows them up beyond the point they are perceived naturally with the objective of emphasizing them (Claridge, 2010). For example, *"I'm tired, I can't lift my hand,"* exaggerates the speaker's exhaustion. Figure 3 shows examples of metaphor and hyperbole.

3 Related Work

Metaphors and hyperboles are the most used figures of speech in everyday utterances (Roger J., 1996). In recent years, significant efforts have been made to understand metaphors and hyperboles, giving rise to interesting techniques to automatically detect and generate them. Troiano et al. (2018) introduced hyperbole detection as a binary classification task, using traditional machine learning algorithms. They also released a dataset named 'HYPO' for hyperbole detection. Kong et al. (2020) introduced 'HYPO-cn', a Chinese dataset for hyperbole detection, and showed that deep learning models can perform better at hyperbole detection with increased data. Biddle et al. (2021) used a BERT (Devlin et al., 2018) based detection system that used the literal sentences of the hyperbolic counterparts to identify the hyperbolic and non-hyperbolic use of words and phrases. They also released a test suite for evaluating models. Tian et al. (2021) proposed a hyperbole generation task. Zhang and Wan (2022) introduced an unsupervised approach for generating hyperbolic sentences from literal sentences and introduced two new datasets 'HYPO-XL' and 'HYPO-L' for their experiments.

Metaphors have been extensively studied even before hyperbole detection was introduced. Tsvetkov et al. (2014) introduced the TSV dataset with 884 metaphorical and non-metaphorical adjective-noun (AN) phrases. They showed that conceptual mapping learnt between literal and metaphorical words is transferable across languages. Mohler et al. (2016) introduced the LCC dataset which contains sentence-level annotations for metaphors in four languages totaling $188,741$ instances. Steen (2010) studied metaphor at the word level and was the first to include function words for metaphor detection with the new VUA dataset. Birke and Sarkar (2006) introduced the TroFi dataset that consists of verbs in their literal and metaphoric form. In recent years, metaphor detection has been explored with the aid of large language models. Choi et al. (2021) used the contextual embeddings from BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) to classify metaphorical sentences. Aghazadeh et al. (2022) probed and analyzed the metaphorical knowledge gained by large language models by testing them on metaphor datasets across languages.

Previous research on metaphor and hyperbole detection typically treats these figurative language forms separately, despite their common properties. In this work, we propose a multi-task approach that simultaneously detects both hyperboles and metaphors, and demonstrate that this approach outperforms individual detection tasks with experimental results and detailed analysis.

4 Task Formulation

For a sentence $\mathbf{x}$ and a corresponding label y or labels $y_1, .., y_k$ $(k > 1)$, we can mathematically formulate the different learning tasks shown in Figure 2 as:

Single Task Learning (STL)

$$y^* = \underset{y \in \{0,1\}}{argmax}\, P(y|\mathbf{x};\theta) \tag{1}$$

$$P(y|\mathbf{x};\theta) = \rho(f(E(\mathbf{x}))) \tag{2}$$

where E and f represent the encoder and the feed-forward neural network (classification head) respectively, θ represents the weights from both E and f, and ρ represents the softmax function. The cross-entropy loss function can be defined as:

$$\mathcal{L} = \frac{-1}{D}\sum_{i=1}^{D}(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)) \tag{3}$$

where D is the number of training samples, y_i and $\hat{y}_i$ are the ith true and predicted labels.

Multi-Task Learning with shared Encoder (MTL-E)

$$y_k^* = \underset{y_k \in \{0,1\}}{argmax}\, P(y_k|\mathbf{x};\theta_k) \tag{4}$$

$$P(y_k|\mathbf{x};\theta_k) = \rho(f_k(E(\mathbf{x}))) \tag{5}$$

where k represents the number of tasks, y_k are the labels, f_k are feed-forward neural networks and θ_k are the weights for the k tasks respectively. For $k = 2$ the loss function can be written as:

$$\mathcal{L} = \lambda\mathcal{L}_1 + (1-\lambda)\mathcal{L}_2 \tag{6}$$

where $\mathcal{L}_1, \mathcal{L}_2$ are task specific losses calculated similar to Eq. 3 and λ is the weighting factor.

Multi-Task Learning with Fully shared layers (MTL-F)

$$y_1^*, y_2^* = \underset{y_1,y_2 \in \{0,1\}}{argmax}\, P(y_1, y_2|\mathbf{x};\theta) \tag{7}$$

Here, the loss is a binary cross-entropy loss:

$$\mathcal{L} = \frac{-1}{D}\sum_{i=1}^{D}\sum_{j=1}^{m}(y_{ij}\log(\sigma(l_{ij}) + (1-y_{ij})\log(1-\sigma(l_{ij})))\quad (8)$$

where σ is the sigmoid function and m is the number of labels and l_{ij} represents the logit value for the ith instance and the jth label.

5 Datasets and Annotations

In this section, we delve into the hyperbole and metaphor datasets used and their annotation details.

5.1 Hyperbole Datasets

Our experiments used two hyperbole datasets: HYPO (Troiano et al., 2018) and HYPO-L (Zhang and Wan, 2022). The HYPO dataset contains 709 hyperbolic sentences each with a corresponding paraphrased literal sentence resulting in $1,418$ sentences. The HYPO-L dataset includes $1,007$ hyperbolic sentences and $2,219$ paraphrased sentences. For each sentence in the HYPO and HYPO-L datasets, we added metaphor labels. Table 1 shows the statistics of the annotated hyperbole datasets.

Dataset (# sentences)	Hyp.	Met.	# sent.
HYPO (1,418)	✓	✓	515
	✓	✗	194
	✗	✓	107
	✗	✗	602
HYPO-L (3,326)	✓	✓	237
	✓	✗	770
	✗	✓	19
	✗	✗	2,200

Table 1: Statistics of annotated hyperbole datasets with metaphor labels, where Hyp. means hyperbole, Met. means metaphor, and #sent is the number of sentences.

5.2 Metaphor Datasets

We used two metaphor datasets: LCC (Mohler et al., 2016) and TroFi (Birke and Sarkar, 2006). We manually annotated $3,838$ (out of $5,482$) sentences in the TroFi dataset and $7,542$ (out of $40,138$) sentences in the LCC dataset with hyperbole labels. For statistics refer to Table 2.

5.3 Annotation Details

We employed four annotators proficient in English in the age group of 24-30. Three annotators were master's students and one had an M.A in linguistics. They were provided with detailed annotation instructions along with examples of hyperbole and metaphors. Each instance of the dataset was annotated once and the annotations were equally divided among the four annotators. We first conducted pilot studies for annotation with randomly sampled 100 sentences from each dataset before proceeding to the final annotation. The Inter Annotator Agreement (IAA) was computed using pairwise Cohen's Kappa score (κ) and Fleiss' Kappa score (K) as reported in Table 3. The IAA between any two annotators is above 0.60 ($0.61 \leq \kappa \leq 0.80$; Cohen (1960)), indicating *substantial* agreement between them. The Fleiss' Kappa score of 0.674 is also considered *substantial* ($0.61 \leq K \leq 0.80$; Landis and Koch (1977)).

Dataset (# sentences)	Met.	Hyp.	# sent.
TroFi (3,838)	✓	✓	209
	✓	✗	1,710
	✗	✓	235
	✗	✗	1,684
LCC (7,542)	✓	✓	615
	✓	✗	3,187
	✗	✓	144
	✗	✗	3,596

Table 2: Statistics of annotated metaphor datasets with hyperbole labels, where Hyp. means hyperbole, Met. means metaphor, and #sent is the number of sentences.

Cohen's Kappa (κ)	A	B	C
B	0.740		
C	0.651	0.653	
D	0.647	0.650	0.707
Fleiss' Kappa (K)			0.674

Table 3: IAA calculations with Fleiss' Kappa and pairwise Cohen's Kappa among the annotators

To ensure the quality of annotations, we randomly sampled 1100 instances with an equal split of hyperbole and metaphor labels across all datasets. The annotators were asked to mark sentences as hyperbole if there was any exaggeration and as metaphors if there were any implicit comparisons. In addition to giving binary labels, we also asked the annotators to mark the part of the sentence that influenced their decisions. Doing this helped us identify any discrepancies in their understanding and correct them. All four annotators received stipends suitable for the tasks.

Task	Model	Hyperbole			Metaphor		
		Precision	Recall	F1	Precision	Recall	F1
STL	$BERT_{lg}$	0.827	0.801	0.811	0.751	0.686	0.711
	$ALBERT_{xxl2}$	0.845	0.871	0.856	0.695	0.736	0.713
	$RoBERTa_{lg}$	**0.883**	0.848	0.864	0.801	0.709	0.745
MTL-F	$BERT_{lg}$	0.853	0.824	0.836	0.799	0.686	0.729
	$ALBERT_{xxl2}$	0.847	0.878	0.860	0.757	**0.761**	0.753
	$RoBERTa_{lg}$	0.879	**0.884**	**0.881***	**0.826**	0.752	**0.787**

Table 4: Comparison of Transformer models using 10-fold cross-validation over three different runs for hyperbole and metaphor detection task on the **HYPO** dataset. Significance test (t-test) p-value ($*$) = 0.0322 (<0.05).

Task	Model	Hyperbole			Metaphor		
		Precision	Recall	F1	Precision	Recall	F1
STL	$BERT_{lg}$	0.670	0.598	0.630	0.561	0.466	0.506
	$ALBERT_{xxl2}$	0.649	0.542	0.589	0.513	0.414	0.456
	$RoBERTa_{lg}$	0.688	0.651	0.667	0.591	0.543	0.563
MTL-F	$BERT_{lg}$	0.655	0.619	0.638	0.552	0.464	0.503
	$ALBERT_{xxl2}$	0.638	0.593	0.614	0.498	0.385	0.430
	$RoBERTa_{lg}$	**0.706**	**0.668**	**0.687***	**0.599**	**0.554**	**0.572**

Table 5: Comparison of Transformer models using 10-fold cross-validation over three different runs for hyperbole and metaphor detection task on the **HYPO-L** dataset. Significance test (t-test) p-value ($*$) = 0.0438 (< 0.05).

6 Experiments

We conduct four experiments: 1) Comparing STL and MTL-F on hyperbole and metaphor datasets, 2) Comparing STL, MTL-E, and MTL-F models, 3) Obtaining sentence-level benchmark results on the metaphor dataset, and 4) Comparing with established baselines for the hyperbole dataset.

For our experiments, we used label-balanced metaphor datasets to address the imbalance caused by fewer hyperbole (Refer to **Appendix** A.2). To ensure a fair comparison, we used mean 10-fold cross-validation obtained over three different runs to compare our models. However, we did not compare our results with existing work on metaphor detection as it does token-level instead of sentence-level metaphor prediction. Finally, we used simple models to highlight the efficacy of a multi-tasked framework for a sophisticated task.

6.1 Hyperbole Baselines

Troiano et al. (2018) used cognitive features, such as imageability, unexpectedness, polarity, subjectivity, and emotional intensity for hyperbole detection, referred to as QQ (i.e. Qualitative and Quantitative). We compare our results with their best-performing Logistic Regression and Naive Bayes models, referred to as **LR+QQ** and **NB+QQ** in Table 9.

Kong et al. (2020) used a combination of the QQ features and a pre-trained BERT, referred to as **$BERT_{base}$+QQ** in Table 9. The QQ features were concatenated with the BERT's output and passed through a linear classifier to predict hyperbole.

Biddle et al. (2021) used literal paraphrases as privileged information and incorporated this information using a triplet loss. We refer to this model as **$BERT_{base}$+PI** in Table 9. We show that our multitask model outperforms all these baselines.

6.2 Experimental Setup

We experiment with `bert-large-uncased` ($BERT_{lg}$) (Devlin et al., 2018), `albert-xxlarge-v2` ($ALBERT_{xxl2}$) (Lan et al., 2020), and `roberta-large` ($RoBERTa_{lg}$) (Liu et al., 2019) models ($h = 16, l = 24$). The best-performing models use the following hyperparameters: For the STL model we use a learning rate of $1e-4$ for 5 epochs and a batch size of 16. For the MTL-E model, the learning rate is $1e-5$ for 20 epochs, a batch size of 32, and the loss weighting factor λ of 0.5 whereas, for the MTL-F model, the learning rate is $1e-5$ for 10 epochs and a batch size of 16. We use Adam (Kingma and Ba, 2015) with eps of $1e-4$ to optimize all our models.

Task	Model	Hyperbole			Metaphor		
		Precision	Recall	F1	Precision	Recall	F1
STL	$BERT_{lg}$	0.557	0.412	0.466	0.531	0.559	0.538
	$ALBERT_{xxl2}$	0.424	0.234	0.294	0.489	0.430	0.454
	$RoBERTa_{lg}$	**0.607**	0.446	0.496	0.542	0.469	0.490
MTL-F	$BERT_{lg}$	0.565	0.433	0.486	0.556	0.525	0.540
	$ALBERT_{xxl2}$	0.487	0.241	0.312	0.516	0.457	0.475
	$RoBERTa_{lg}$	0.605	**0.529**	**0.561**	**0.565**	**0.587**	**0.573***

Table 6: Comparison of Transformer models using 10-fold cross-validation accuracy over three different runs for hyperbole and metaphor detection on the label balanced **TroFi** dataset. Significance test (t-test) p-value ($*$) < 0.0001.

Task	Model	Hyperbole			Metaphor		
		Precision	Recall	F1	Precision	Recall	F1
STL	$BERT_{lg}$	0.649	0.542	0.589	0.758	0.736	0.745
	$ALBERT_{xxl2}$	0.591	0.546	0.564	0.723	0.757	0.739
	$RoBERTa_{lg}$	**0.692**	0.604	0.645	**0.802**	0.787	0.794
MTL-F	$BERT_{lg}$	0.633	0.531	0.575	0.750	0.774	0.760
	$ALBERT_{xxl2}$	0.614	0.425	0.499	0.709	0.785	0.744
	$RoBERTa_{lg}$	0.630	**0.691**	**0.659**	0.798	**0.812**	**0.805***

Table 7: Comparison of Transformer models using 10-fold cross-validation over three different runs for hyperbole and metaphor detection on the label balanced **LCC**. Significance test (t-test) p-value ($*$) = 0.0221 (< 0.05).

6.3 Hyperparameter Details

We did hyperparameter search manually with the following search space: number of epochs = [5, 7, 10, 15, 20, 25], learning rate = [1e-5, 5e-5, 1e-4, 2e-4, 5e-4], and batch size = [4, 8, 16, 32, 64].

The hyperparameters of the best-performing models have been mentioned in Section 6.2. The training runs for STL, MTL-E, and MTL-F models were 150, 600, and 300 respectively and 30 evaluation runs each.

6.4 Hypothesis Testing

We used t-test, which is a statistical test used to determine if there is a significant difference between the means of two groups. The p-value here is a statistical measure that is used to assess the evidence against a null hypothesis. A p-value < 0.05 is typically considered to be statistically significant. The null hypothesis to reject here is that both the samples for STL and MTL-F models come from the same distribution.

For all our experiments, we obtain a p-value < 0.05 indicating that the samples are indeed coming from different distributions. This shows that the improvement obtained by the MTL-F model over the STL model is statistically significant.

7 Results

STL vs. MTL-F models We use identical experimental setups to compare the results obtained from the STL and MTL-F approach on all four datasets.

1. HYPO results: The comparative analysis results for the HYPO dataset are in Table 4. For all the models we observe that the MTL-F performs better than the corresponding STL. Overall the $RoBERTa_{lg}$ MTL-F model achieves the best recall of 0.884 and F1 of **0.881** (1.96% ↑) for hyperbole detection and a p-value of **0.0322**.

2. HYPO-L results: The comparative analysis results for the HYPO-L dataset are in Table 5. For all the models we observe that the MTL-F performs better than the corresponding STL for hyperbole detection. Overall the $RoBERTa_{lg}$ MTL-F model achieves the best precision of 0.706, recall of 0.668, and F1 of **0.687** (2.99% ↑) for hyperbole detection and a p-value of **0.0438**.

3. TroFi results: The comparative analysis results for the label-balanced TroFi dataset is in Table 6. For all the models we observe that the MTL-F performs better than the corresponding STL for metaphor detection. Overall the $RoBERTa_{lg}$ MTL-F model achieves the best precision of 0.565, recall of 0.587, and F1 of **0.573** (**16.93%** ↑) for metaphor

Task	Model	Hyperbole			Metaphor		
		Precision	Recall	F1	Precision	Recall	F1
STL	RoBERTa$_{lg}$	**0.883**	0.848	0.864	**0.802**	0.787	0.794
MTL-E	RoBERTa$_{lg}$	0.859	0.878	0.867	0.792	0.808	0.799
MTL-F	RoBERTa$_{lg}$	0.879	**0.884**	**0.881**	0.798	**0.812**	**0.805**

Table 8: Comparison of STL, MTL-E and MTL-F models using 10-fold cross-validation over three different runs on the **HYPO** dataset for hyperbole detection and the label balanced **LCC** dataset for metaphor detection. The metaphor column gives the **benchmark** results (sentence-level) on the label-balanced LCC dataset.

Figure 4: Examples show the improvement in the focus of the MTL-F model over the STL model for two cases: a) Classifying a hyperbolic sentence in the presence of metaphor labels. b) Classifying a metaphoric sentence in the presence of hyperbole labels. (Darker colors indicate higher attention)

	Model	P	R	F1
Baselines	LR+QQ	0.679	0.745	0.710
	NB+QQ	0.689	0.696	0.693
	BERT$_{base}$	0.711	0.735	0.709
	BERT$_{base}$+QQ	0.650	0.765	0.671
	BERT$_{base}$+PI	0.754	0.814	0.781
Ours	RoBERTa$_{lg}$ STL	**0.883**	0.848	0.864
	RoBERTa$_{lg}$ MTL-E	0.859	0.878	0.867
	RoBERTa$_{lg}$ MTL-F	0.879	**0.884**	**0.881**

Table 9: HYPO Results. Precision (P), recall (R) and F1 score for baseline models compared to our work.

detection and a p-value < **0.0001**.

4. LCC results: The comparative analysis results for the label-balanced LCC dataset are in Table 7. For all the models we observe that the MTL-F performs better than the corresponding STL for metaphor detection. Overall the RoBERTa$_{lg}$ MTL-F model achieves the best recall of 0.812, and F1 of **0.805** (1.38% ↑) for metaphor detection and a p-value of **0.0221**.

We observe: a) The MTL-F model helps in achieving generalization under the presence of both hyperbole and metaphor labels. b) The p-values (30 samples) suggest that the MTL-F results are statistically significant over the STL results with 95% confidence for all the datasets (**Appendix** 6.4).

STL vs. MTL-E vs. MTL-F models Table 8 reports the comparison of these three models on the HYPO and LCC datasets for hyperbole and metaphor detection respectively. We observe that, in comparison to the STL model, the MTL-E model performs better in general whereas the MTL-F model performs significantly better, achieving the best F1 score of **0.881** and **0.805** on the HYPO and LCC datasets respectively. (See **Appendix** A.3).

Benchmark Results We report the benchmark results for sentence-level detection on the label balanced LCC dataset in Table 8 (check the Metaphor column). Our RoBERTa$_{lg}$ MTL-F model achieves the best recall of **0.812** and F1 of **0.805**.

Baseline Comparison Table 9 reports the comparison of our work with baseline models on the HYPO dataset for hyperbole detection. Our RoBERTa$_{lg}$ MTL-F model achieves the best recall of **0.884** (8.59% ↑) and F1 of **0.881** (**12.03**% ↑) as compared to the recall of 0.814 and F1 of 0.781 of the state-of-the-art system.

8 Analysis

We divide our analysis into two subsections: 1) A comparison of the STL and MTL-F models, and 2) Error analysis of the MTL-F model.

Sentences	Actual	MTL-F	STL	
			HD	MD
Your plan is too risky, it's a suicide.	H, M	H, M	NH	NM
I'm not staying here any longer!	NH, NM	NH, NM	H	NM
This kind of anger rages like a sea in a storm.	H, NM	H, NM	H	M
My ex boyfriend! Treacherous person!	NH, NM	NH, NM	H	M
They cooked a turkey the size of a cow.	H, M	H, M	H	NM
Her strength awoke in poets an abiding love.	NH, M	NH, M	H	M
My sister is a vortex of intelligence in space.	H, M	H, M	H	M
The act of love strongly resembles severe pain.	NH, NM	NH, NM	NH	NM

Table 10: Some cases where the MTL-F performs better than the STL for hyperbole detection (HD) and metaphor detection (MD). Here H denotes a hyperbolic sentence, M denotes a metaphoric sentence, NH denotes a non-hyperbolic sentence, and NM denotes a non-metaphoric sentence. Notations in red indicate incorrect detection.

Sentences	Actual	MTL-F
What kind of sorcery is this?	H, M	NH, NM
You're grumpy.	NH, NM	NH, M
..this car is more a sophisticated piece of machinery than a regular car.	NH, NM	H, NM
Stop bothering him: you're inviting trouble.	NH, M	NH, NM
The work of the farm seemed to rest entirely on this horse's mighty shoulders.	H, M	NH, M

Table 11: Error cases where MTL-F fails in the detection task. Here H denotes a hyperbolic sentence, M denotes a metaphoric sentence, NH denotes a non-hyperbolic sentence, and NM denotes a non-metaphoric sentence.

8.1 Comparative Analysis

Under similar experimental setups, we compare the STL and MTL-F models on example sentences obtained from the different test sets of the cross-validation run of the HYPO dataset as shown in Table 10. We consider the following 4 cases:

1. Hyperbolic and Metaphoric: *"They cooked a turkey the size of a cow,"* is both hyperbolic and metaphorical. Here, the exaggeration is evident as the size of the turkey is being compared to that of a cow, which allows both the STL and MTL-F models to make correct hyperbole predictions. However, for metaphor prediction, the MTL-F model correctly identifies the implicit meaning of "size being big" under the influence of the correct hyperbole label, while the STL model fails to do so.

Next, for the example sentence, *"Your plan is too risky, it's a suicide,"* the exaggeration and the metaphoricity are very intricate. The words *risky* and *suicide* make it difficult for the STL model to detect the labels, but the MTL-F model accurately identifies them. This can be attributed to the MTL-F model's ability to learn from both labels.

2. Non-Hyperbolic and Non-Metaphoric: In some cases, the STL model may incorrectly classify sentences that are non-hyperbolic and non-metaphoric due to ambiguous language. For example, in the sentence *"I'm not staying here any longer!"* the words *staying* and *longer* may give the impression of exaggeration, causing the STL model to incorrectly classify it as hyperbolic.

However, the MTL-F model, by learning both hyperbole and metaphor detection simultaneously, is able to identify such cases as non-hyperbolic. Similarly, in *"My ex boyfriend! Treacherous person!"* the word *treacherous* may lead the STL model to incorrectly classify it as hyperbolic and metaphoric, but the MTL-F model classifies it correctly.

3. Hyperbolic and Non-Metaphoric: For this category, we notice that similes can cause confusion. For instance, in the sentence *"This kind of anger rages like a sea in a storm," anger* is explicitly compared to *sea in a storm* through the word *like*. The MTL-F model is able to distinguish this as a simile, whereas the STL model fails to do so.

4. Non-Hyperbolic and Metaphoric: Here we observe that the use of figurative language is subtle. For instance, in *"Her strength awoke in poets an abiding love," awoke* is used metaphorically, which is correctly identified by both the STL and MTL-F models. However, the STL model incorrectly tags it as hyperbolic, while the MTL-F model learns to identify such sentences as non-hyperbolic.

Analysis of attention weights:
Additionally, we also examine the attention weights from the final layer to gain an insight into the performance of the MTL-F model compared to the STL model. We use the weights associated with the **[CLS] / <s>** ([CLS] for BERT and <s> for RoBERTa) token normalized over all the attention heads.

First, we compare the STL and MTL-F models for the task of hyperbole detection. Figure 4. shows attention weight comparison of example sentences. For the sentence *"Hope deferred makes the heart sick,"* we observe that the MTL-F model focuses on the words *heart* and *sick* that indicate exaggeration, while the STL model focuses on other irrelevant words. Similarly, for *"Books are food for avid readers,"* the MTL-F model correctly focuses on the words *Books*, *food* and *readers*. This suggests that the MTL-F model is better at paying attention to relevant words in the sentence due to its knowledge of both hyperbole and metaphor detection.

Next, for metaphor detection, the presence of hyperbole labels during training helps the MTL-F to learn to correctly attend to relevant tokens. For example, in *"After workout I feel I could lift a sumo wrestler,"* the MTL-F focuses on the words *lift* and *wrestler* to correctly identify it as metaphoric. Similarly, for *"Seeing my best friend again would mean the world to me,"* the MTL-F pays the maximum attention to the words *would*, *mean*, and *world* which is the reason for metaphoricity here.

8.2 Error Analysis

We also analyzed the misclassifications for the MTL-F model, some of which have been included in Table 11. We observe that the primary reason for misclassifications in the MTL-F model is the lack of context in identifying the exaggeration or metaphoricity. For instance, *"What kind of sorcery is this?"* is a commonly used figurative sentence but the absence of any context makes it difficult for the MTL-F model to classify it correctly as both hyperbolic and metaphoric.

Next, we found cases such as *"You're grumpy,"* where the MTL-F model tags them incorrectly as metaphoric. Such mistakes could be attributed to the model learning to identify implicit comparisons but failing to identify that *grumpy* here is an attribute not a comparison.

9 Conclusion and Future work

We have presented a novel multi-tasking approach to the detection of hyperboles and metaphors. We augmented the annotations of two hyperbole datasets with metaphor labels and that of two metaphor datasets with hyperbole labels. This allowed multi-task learning of metaphor and hyperbole detection, which outperforms single-task learning on both tasks. We establish a new SOTA for hyperbole detection and a new benchmark for sentence-level metaphor detection. The take-away message is that metaphor and hyperbole detection help each other and should be done together.

We plan to extend our framework of exploiting linguistic relatedness and thereby creating MTL detection systems, to all forms of figurative languages like proverbs, idioms, humour, similes, and so on.

10 Limitations

The scope of this work is limited to sentence-level detection due to the absence of any span-level annotated datasets for hyperbole detection. Also, we could only partially annotate the metaphor datasets due to resource constraints. Finally, we did not try sophisticated large language models in our work as our goal was to demonstrate the effectiveness of multitasking using a simple model, rather than to test the performance of more sophisticated models.

11 Ethical Considerations

We perform our experiments on existing hyperbole and metaphor datasets by adding additional labels to them. Some of the examples in these datasets use slurs, abuses, and other derogatory terms to bring out exaggeration or implicit comparison. Our models may also propagate these unintended biases due to the nature of the datasets. We urge the research community to use our models and these datasets with caution and we are fully committed to removing discrepancies in the existing hyperbole and metaphor datasets in the future.

Acknowledgements

We thank our annotation team for all the efforts taken to ensure quality during the annotation. Special mention to Akash Bhagat from the annotation team for his efforts. We thank the anonymous reviewers for their insightful feedback. We also thank our colleagues at CFILT Lab, IIT Bombay for their valuable comments and suggestions.

References

Muhammad Abulaish, Ashraf Kamal, and Mohammed J. Zaki. 2020. A survey of figurative language and its computational detection in online social networks. *ACM Trans. Web*, 14(1).

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages.

Rhys Biddle, Maciek Rybinski, Qian Li, Cecile Paris, and Guandong Xu. 2021. Harnessing privileged information for hyperbole detection. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 58–67, Online. Australasian Language Technology Association.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336, Trento, Italy. Association for Computational Linguistics.

Robyn Carston and Catherine Wearing. 2015. Hyperbolic language and its relation to metaphor and irony. *Journal of Pragmatics*, 79:79–92.

Minjin Choi, Sunkyung Lee, Eun-Kyu Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. *ArXiv*, abs/2104.13615.

Claudia Claridge. 2010. *Hyperbole in English: A corpus-based study of exaggeration*. Cambridge University Press.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Li Kong, Chuanyi Li, Jidong Ge, Bin Luo, and Vincent Ng. 2020. Identifying exaggerated language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7024–7034, Online. Association for Computational Linguistics.

George Lakoff. 1993. The contemporary theory of metaphor.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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*, abs/1907.11692.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Comput. Surv.*, 53(2).

Richard M. Roberts and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological Science*, 5(3):159–163.

Kreuz Roger J. 1996. *Figurative language occurrence and co-occurrence in contemporary literature*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

D. Sperber and D. Wilson. 2008. Relevance: Communication and cognition. *A Deflationary Account of Metaphor*, page 84 – 108. Cited by: 1.

Gerard Steen. 2010. A method for linguistic metaphor identification : from mip to mipvu.

Yufei Tian, Arvind krishna Sridhar, and Nanyun Peng. 2021. HypoGen: Hyperbole generation with commonsense and counterfactual knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1583–1593, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Yunxiang Zhang and Xiaojun Wan. 2022. MOVER: Mask, over-generate and rank for hyperbole generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6018–6030, Seattle, United States. Association for Computational Linguistics.

A Appendix

A.1 Experimental Details

For experiments, we use the NVIDIA A100-SXM4-80GB GPU. Check Table 12 for further details.

Model	#Parameters	Run time
BERT_{lg}	$\sim 335M$	$\sim$ 25 mins
$\text{ALBERT}_{\text{xxl2}}$	$\sim 222M$	$\sim$ 45 mins
$\text{RoBERTa}_{\text{lg}}$	$\sim 355M$	$\sim$ 26 mins

Table 12: Additional details of the models along with their number of parameters and run time.

A.2 Label Balanced Metaphor Datasets

As discussed in Section 6, we used label-balanced metaphor datasets to address the imbalance caused by fewer hyperbole. Table 13 reports the statistics of the label-balanced metaphor datasets.

Dataset	#H	#NH	#M	#NM
TroFi	444	1100	709	835
LCC	634	1400	1217	817

Table 13: Statistics of label balanced metaphor datasets. #H, #NH, #M, and #NM represent the number of hyperboles, non-hyperboles, metaphors, and non-metaphors respectively.

A.3 STL vs. MTL-E vs. MTL-F models

Detailed comparison of the STL, MTL-E and MTL-F models are reported in Table 14 and Table 15. For hyperbole detection we used the HYPO dataset whereas for metaphor detection we used label balanced LCC dataset. Table 8 in the paper reports the comparison of only the best performing models for brevity.

Task	Model	Hyperbole		
		Precision	Recall	F1
STL	BERT_{lg}	0.827	0.801	0.811
	$\text{ALBERT}_{\text{xxl2}}$	0.845	0.871	0.856
	$\text{RoBERTa}_{\text{lg}}$	**0.883**	0.848	0.864
MTL-E	BERT_{lg}	0.817	0.788	0.802
	$\text{ALBERT}_{\text{xxl2}}$	0.726	0.746	0.734
	$\text{RoBERTa}_{\text{lg}}$	0.859	0.878	0.867
MTL-F	BERT_{lg}	0.853	0.824	0.836
	$\text{ALBERT}_{\text{xxl2}}$	0.847	0.878	0.860
	$\text{RoBERTa}_{\text{lg}}$	0.879	**0.884**	**0.881**

Table 14: Comparison of STL, MTL-E, and MTL-F models on the HYPO dataset for hyperbole detection.

Task	Model	Metaphor		
		Precision	Recall	F1
STL	BERT_{lg}	0.758	0.736	0.745
	$\text{ALBERT}_{\text{xxl2}}$	0.723	0.757	0.739
	$\text{RoBERTa}_{\text{lg}}$	**0.802**	0.787	0.794
MTL-E	BERT_{lg}	0.737	0.762	0.749
	$\text{ALBERT}_{\text{xxl2}}$	0.654	0.610	0.629
	$\text{RoBERTa}_{\text{lg}}$	0.792	0.808	0.799
MTL-F	BERT_{lg}	0.750	0.774	0.760
	$\text{ALBERT}_{\text{xxl2}}$	0.709	0.785	0.744
	$\text{RoBERTa}_{\text{lg}}$	0.798	**0.812**	**0.805**

Table 15: Comparison of STL, MTL-E, and MTL-F on the label balanced LCC dataset for metaphor detection.

B Case study: Hyperbole and Metaphor Detection using LLMs

We tested the zero-shot capabilities of recent LLMs for hyperbole and metaphor detection. We experimented with a varied set of prompts on the open sources LLMs- LLaMA (Touvron et al., 2023) and BLOOM (Scao et al., 2022). We found that they do not have any deeper understanding of hyperboles and metaphors and performed poorly in the classification task. Our observations with ChatGPT are as follows. ChatGPT had a good understanding of what hyperbole and metaphor are and was able to classify popular examples correctly. However, it struggled to perform multi-class classification when sentences had both hyperbole and metaphor properties. It also struggled to identify many hyperboles. When pointed out explicitly that it has made a mistake, and a particular part of the sentence is an indication of hyperbole or metaphor, it was able to correct itself and explain the reason. We experienced such issues with all combinations of presence or absence of metaphor and hyperbole. The example instances demonstrating the same are shown below.

B.1 Example 1

User*: Can you detect if the sentence has hyperbole or metaphor or both in it? Sentence: That sharp bend is a death-trap for motorists.*
ChatGPT*: Yes, I can detect that the sentence contains a metaphor. The sentence "That sharp bend is a death-trap for motorists" is a metaphor because it is comparing the sharp bend to a death-trap, suggesting that the bend is extremely dangerous and potentially deadly for motorists. The metaphor is used to create a vivid and impactful image in the reader's mind.*
User*: Does it have a hyperbole?*
ChatGPT*: No, the sentence does not contain hyperbole. Hyperbole is an exaggerated statement or claim not meant to be taken literally. The sentence "That sharp bend is a death-trap for motorists" is not an exaggeration, but rather a metaphorical comparison between the sharp bend and a death-trap.*
User*: Isn't "death-trap" a hyperbole?*
ChatGPT*: You are correct, "death-trap" in this context can be considered a hyperbole.....*

B.2 Example 2

User*: Can you detect if the sentence has hyperbole or metaphor in it? Sentence: Retailers, meanwhile, are trying out a number of ways to target these shoppers.*
ChatGPT*: The sentence "Retailers, meanwhile, are trying out a number of ways to target these shoppers" does not contain hyperbole or metaphor. It is a straightforward statement conveying information about the actions of retailers in targeting a particular group of shoppers.*
User*: Isn't the word "target" a metaphor here?*
ChatGPT*: You are correct that "target" in this sentence can be considered a metaphor. In this context, "target" is being used figuratively to mean that retailers are trying to aim their marketing efforts towards a particular group of shoppers. The word "target" is being used in a non-literal sense to describe the retailers' actions. Therefore, the sentence does contain a metaphor. Thank you for pointing that out.*

It can be seen that in both examples, the model initially makes the wrong assumption about the sentence being a hyperbole or metaphor. It was able to correct itself only after bringing attention to the important word in the sentence. We have shown that the correct words get more attention through our multi-tasked approach indicating the reason for better detection accuracy.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 10

☑ A2. Did you discuss any potential risks of your work?
Section 11

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Sections 5 and 6

☑ B1. Did you cite the creators of artifacts you used?
Section 5 and Section 6.2

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 5

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5

C ☑ Did you run computational experiments?

Section 6.2

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section A.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 6.2 and Section 6.3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 7

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 6

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 5.3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 5.3

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 5.3

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.