

SIcBERT: Scalar Implicature classification with DeBERTa

Vittorio Ciccarelli | 3118588
Vittorio.Ciccarelli@hhu.de

Abstract

Scalar implicatures are crucial for natural language understanding, influencing how meaning is derived beyond literal interpretations. This study evaluates the performance of the DeBERTa model using a dataset tailored for the task of Scalar Implicature classification. To achieve this, we fine-tuned DeBERTa using the SIGA (Scalar Implicatures with Gradable Adjectives) dataset from (Nizamani et al., 2024) before evaluating its performance on our dataset. We then compared results against a non-fine-tuned version of the model. The results, measured using F1 scores, indicate that while fine-tuning slightly improved the model’s ability to derive scalar implicatures with gradable adjectives, the model fails with a dataset that contains scalar expressions other than gradable adjectives, exhibiting a bias towards the neutral class. These findings highlight persistent challenges in the computational modeling of pragmatic inferences and call for further experimentation on the task of inferring scalar implicatures.

Keywords: Scalar Implicatures, Classification, Large Language Models, Natural Language Inferencing, DeBERTa.

1 Introduction

One of the core components of human communication is the ability to go beyond what is “explicitly said” and understand implicated meanings through the enrichment of utterances with pragmatic implicatures (Grice, 1975).

Although humans generally perform this task unconsciously, understanding non-literal meanings has long been considered a difficult task for language models (Yue et al., 2024). To this end, the exponential progress in artificial intelligence (AI) and in particular in the field of Natural Language Processing (NLP), has sparked a heated debate about whether Large

Language Models (LLMs) and LLM-driven chatbots, such as ChatGPT, truly comprehend language in a human-like manner or merely recognize statistical patterns (Chomsky et al., 2023; Piantadosi, 2023). Over the past years, research areas such as Natural Language Inferencing (NLI) were established to test these model’s ability to infer pragmatic meanings: in this task, a model is tested with sentence pairs for which it has to decide whether there is an entailment, contradiction or a neutral relation (Jeretic et al., 2020). While LLMs have proven to be successful in various linguistic tasks, their actual performance when handling non-literal meaning remains to be explored, especially on pragmatic inferencing tasks. A particular type of pragmatic inference are Scalar Implicatures (SIs). Scalar implicatures are inferences that result from the comparison between an uttered sentence and a slightly different alternative sentence that hasn’t been uttered (Chemla and Bott, 2014).

Given the increasing importance of evaluating LLMs’ pragmatic understanding, this study investigates DeBERTa’s (He et al., 2020) ability to infer scalar implicatures by fine-tuning the model utilizing dataset from the (Nizamani et al., 2024) paper, which focuses on scalar and gradable adjectives (Kennedy and McNally, 2005). We then assess its performance on a separate dataset consisting of sentence pairs in which different types of scalar expressions are present, comparing F1 scores from both the fine-tuned and non-fine-tuned versions of the DeBERTa model. Our findings reveal that fine-tuning improved DeBERTa’s performance on the SIGA dataset, as previously observed by the authors of the paper themselves.

However, when applied to our dataset, both the fine-tuned and non-fine-tuned models predominantly predicted the neutral class. The fine-tuned model assigned a small probability to contradiction, while the non-fine-tuned model did not. Neither model predicted entailment for our dataset. As we will see, differences in dataset format, particularly the inclusion of "but not" statements, may have affected the model's performance on our dataset. Thus, these results underline persistent challenges in the computational modeling of pragmatic inferences through LLMs.

2 Theoretical Foundations

2.1 Scalar Implicatures

The literal meaning of a sentence can be enriched by taking into account extra-linguistic information, such as general rules of communication and social interaction, information about the context of the utterance or the assumed knowledge between speaker and addressee (Chemla and Bott, 2014). According to Grice (1975), the implementation of these pragmatic procedures leads to the formation of so-called conversational implicatures. Implicatures deal precisely not with what is being explicitly said (literal meaning), but with what is being 'implied'. A particular type of implicatures are the so-called 'Scalar Implicatures' (SI). Scalar Implicatures are inferences that result from the comparison between an uttered sentence and a slightly different alternative sentence that hasn't been uttered (Chemla and Bott, 2014). Consider as an example the sentence:

- (1) John read some of the books.

The unused alternative to such sentence, for example, would have been:

- (2) John read all the books.

Upon hearing (1), through pragmatic reasoning, the listener notices that (2), although more informative, was not uttered, therefore the listener is led to believe that their interlocutor did not believe that (2) was true, which gives rise to the Scalar Implicature:

- (3) John didn't read all the books.

2.2 Large Language Models and SIs

In recent years, several studies have focused on the topic of Scalar Implicatures in the context of Large Language Models.

Nizamani et al. (2024) for example, which forms the basis of the rest of this research, proposed an NLI dataset with the specific aim of testing the models' ability to interpret utterances containing scalar implicatures arising from the presence of gradable adjectives (Kennedy and McNally, 2005), demonstrating that using their dataset for fine-tuning DeBERTa led to an improvement in the model's ability to derive scalar implicatures, but nonetheless revealing a poor performance when compared to other standard NLI tasks. In Qiu et al. (2023), on the other hand, the authors tested through a series of experiments whether ChatGPT was able to enrich literal meaning with pragmatic implicatures, coming to the conclusion that the model follows deterministic patterns. Jeretic et al. (2020) tested whether LLMs are capable of inferring both (scalar) implicatures and presuppositions, finding that BERT (Devlin et al., 2019) reliably treated scalar implicatures associated with determiners (e.g., 'some' vs. 'all') as entailments, but highlighting that LLMs often fail to derive pragmatic inferences that extend beyond the "some/all" triggers. Finally, Lipkin et al. (2023), examined LLMs ability to infer pragmatic meanings in sentences in which gradable adjectives such as 'strong' are responsible for scalar implicatures, coming to the conclusion that, although LLMs exhibit emerging pragmatic capabilities, they remain limited and unreliable in more complex contexts.

In summary, while contemporary LLMs demonstrate some competence in pragmatic inference, significant challenges still persist, particularly when it comes to handling more complex inferences.

3 Data

The present research utilizes two datasets: the SIGA (Nizamani et al., 2024) dataset, which was used to fine-tune the DeBERTa model, and a modified version of the dataset by Sun et al. (2024), originally developed to investigate how different scalar expressions give rise to scalar implicatures (SIs) at varying rates. The latter was constructed from naturally occurring Twitter data and involved a paraphrase similarity task, and was used here to evaluate the performance of the fine-tuned model.

3.1 SIGA Dataset

The SIGA dataset comprises sentence pairs containing two scalar adjectives separated by “but not”, leveraging the reinforceability principle of scalar implicatures (Hirschberg, 1985) along with textual context. The sentences were mined through a predefined list of 88 adjective pairs from de De Melo and Bansal (2013). The resulting dataset consisted of 1,600 selected examples.

3.2 Scalar Diversity Corpus

Once the model was fine-tuned with the SIGA dataset, we evaluated its performance using a tailored version of Sun et al. (2024) dataset, which contains sentences with scalar expressions (quantifiers, adverbs, modals, and adjectives) collected from Twitter via the Twitter API (USA). To assess whether the scalar implicature was derived, participants were asked to rate the sentence pairs using a Likert scale from 1 (very different meaning) to 7 (same meaning).

We later used these similarity ratings to establish gold labels for our classification task.

4 Methodology

We followed the methodology outlined in Figure 1.

4.1 Dataset Preparation

The dataset we crafted from Sun et al. (2024) consisted of sentence pairs, including entailment, contradiction, and neutral examples. We generated 500 neutral examples by randomly

combining pairs of sentences from the dataset in order to form sentences in a neutral relationship. The dataset was created according to the following steps:

- **Entailment and Contradiction Sentences:** the gold labels for Entailment (2) and Contradiction (0) were created by calculating the mean similarity score for each sentence pair. If the mean similarity score was 4 or above, the sentence pair was labeled as Entailment; if it was below 4, it was labeled as Contradiction.
- **Neutral Sentence Generation:** neutral samples were generated by randomly pairing sentences from the dataset, ensuring that each pair was unique and did not repeat any existing combination. These pairs were not logically entailed or contradictory, preserving the neutral relationship between premise and hypothesis, as in the following example:

- (4) Premise: “The food was warm.”
 Hypothesis: “The movie was great.”

- **Text Cleaning and Formatting:** the dataset underwent some processing, which included (1) the removal of tokens from the text and conversion to lowercase, (2) elimination of unnecessary metadata and participant information columns, (3) calculation of the mean similarity score for each sentence pair, with gold labels added to a new column, and lastly (4) format adjustments to make the dataset compatible with the SIGA model (renaming columns, etc.). The final dataset consisted of 1894 pairs of sentences:

Label	Count
Contradiction	441
Neutral	500
Entailment	953

Table 1: Label distribution for our dataset

4.2 Fine-Tuning Process

Once we obtained the evaluation dataset, we fine-tuned DeBERTa (He et al., 2020). The fine-tuning procedure followed the guidelines provided in the SIGA GitHub repository,

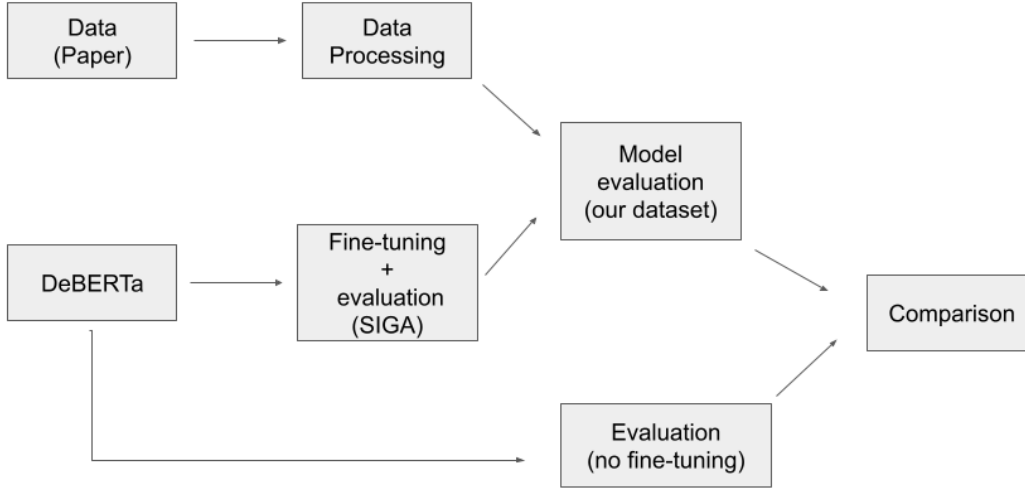


Figure 1: Pipeline of the project

where the model was trained on their dataset. The fine-tuning process involved the following steps: (1) we initialized the pre-trained SIGA model as described in the SIGA repository, (2) the SIGA dataset was used as the training data to fine-tune the model on the task of classifying sentence pairs into the three categories (Contradiction, Neutral, and Entailment), (3) the fine-tuned model’s performance was evaluated using F1 scores for each class on both in-domain and out-of-domain examples from the SIGA dataset.

4.3 Model Evaluation

Once the fine-tuning process was over, we evaluated the model on our dataset.¹ As a baseline for comparison, we used a non-fine-tuned version of DeBERTa which had not undergone any additional training on our dataset. For both models, F1 scores were computed for each of the three classes (Contradiction, Neutral, and

Entailment).

5 Results

5.1 Evaluation on SIGA Dataset (Fine-Tuned Model)

The fine-tuned model performed well on the SIGA dataset, especially in out-of-domain testing, which comprised examples that were not included in the fine-tuning process, where for the entailment label the model achieved an F1 score of 0.6435 (see table 2 for full results).

5.2 Evaluation on Our Dataset (Fine-Tuned Model)

We then evaluated the fine-tuned model on our dataset, which contained both the entailment, contradiction, and the 500 neutral examples that we generated. The fine-tuned model exhibited a limited ability to predict the neutral class, however with a low F1 score of 0.1830. While a small proportion of predictions fell under contradiction (F1 score of 0.0748), the model did not predict any entailment instances (F1 score of 0).

¹To ensure compatibility with our dataset, several adjustments were made to the original Python scripts from the SIGA GitHub repository: (1) File paths were updated to point to our local dataset, (2) Variable names were modified to ensure that the dataset was correctly loaded and processed for evaluation, (3) These changes ensured that the model could be tested on our dataset in the same way it was tested on the SIGA dataset. All these changes to the original scripts can be found on the [Github repository](#) for this project.

Table 2: F1 Scores for the Models on Different Datasets

Test Data	Model	F1 Score		
		C	N	E
SIGA (In-Domain)	Fine-Tuned	0.5851	0.0000	0.4806
SIGA (Out-of-Domain)	Fine-Tuned	0.5137	0.1569	0.6435
Our Dataset	Fine-Tuned	0.0748	0.1830	0.0000
Our Dataset	Non-Fine-Tuned	0.0000	0.3000	0.0000

5.3 Evaluation on Our Dataset (Non-Fine-Tuned Model)

Lastly, we tested a non-fine-tuned version of the model on our dataset for comparison: similar to the fine-tuned model, the non-fine-tuned model leaned toward the neutral class (F1 score of 0.3000), and did not predict any contradiction or entailment instances.

5.4 Performance Comparison

From the comparison of the fine-tuned and non-fine-tuned models’ performances on our dataset, we can make the following observations: (1) both models predicted the neutral class, with F1 scores of 0.1830 (fine-tuned) and 0.3000 (non-fine-tuned), (2) the fine-tuned model assigned a small number of predictions to the contradiction class (F1 score of 0.0748), while the non-fine-tuned model did not predict any contradiction label at all, (3) neither the fine-tuned nor the non-fine-tuned models predicted entailment for any sentence pairs in our dataset. In general, both models favored the neutral class, but with a drop in performance for the fine-tuned model. In addition to that, only the fine-tuned model predicted contradiction, albeit with a very low F1 score.

5.5 Dataset Differences and Impact on Model Performance

The primary difference between our dataset and the SIGA dataset lies in the structure of the hypotheses. The SIGA dataset includes sentence pairs with gradable adjectives and no contrastive elements:

- (5) Premise: “The food was good.”
Hypothesis: “The food was great.”

while our dataset includes sentences containing “but not” as a contrastive statement in each sentence:

- (6) Premise: “Be wary of the self-righteous and the idealistic.”
Hypothesis: “Be wary, but not scared of the self-righteous and the idealistic.”

This format difference may have influenced the model’s ability to effectively classify sentence pairs into the appropriate categories. Furthermore, the SIGA dataset only contained sentence pairs with gradable adjectives, whereas our data included, along with gradable adjectives, other scalar items such as adverbs, quantifiers and modals, which were not included in the model fine-tuning and could have thus affected the model’s overall performance. Interestingly, the fine-tuned model performed worse than the non-fine-tuned version when classifying neutral examples in our dataset. At this stage, we do not have a clear explanation for this drop in performance and we believe further investigation is needed to understand the underlying reasons for this behavior.

6 Conclusion and Future Directions

The present research aimed at evaluating DeBERTa’s ability to infer scalar implicatures. After fine-tuning the model with the SIGA dataset, which consisted of sentence pairs containing SIs stemming from gradable adjectives, the model was evaluated against a second dataset containing a vast variety of scalar items. According to our findings, fine-tuning improved model performance on the SIGA dataset, especially on out-of-domain examples, for which the entailment label reached a F1 score of 0.6435. On the other hand, the model

struggled when applied to our dataset, both in its fine-tuned and non-fine-tuned version, exhibiting in both cases a bias toward the neutral class, with very few predictions for contradiction and none for entailment. We believe that the format differences in our data, such as the inclusion of contrastive statements (WEAK, but not STRONG), played an important role in the model’s ability to make accurate predictions. On the same line, the presence of different scalar items in the evaluation dataset might have also impacted the performance of the model, which was instead only fine-tuned for scalar implicatures with gradable adjectives and might have thus struggled to generalize to different types of scalar expressions. Therefore, we believe that further experimentation, such as fine-tuning the model with a more varied dataset and a broader set of scalar expressions, is essential for achieving better results.

References

- Emmanuel Chemla and Lewis Bott. 2014. Processing inferences at the semantics/pragmatics frontier: Disjunctions and free choice. *Cognition*, 130(3):380–396.
- Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. Noam chomsky: The false promise of chatgpt. *The New York Times*, 8.
- Gerard De Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.
- Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Julia Bell Hirschberg. 1985. *A theory of scalar implicature (natural languages, pragmatics, inference)*. Ph.D. thesis, University of Pennsylvania.
- Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models impressive? learning implicature and presupposition. *arXiv preprint arXiv:2004.03066*.
- Christopher Kennedy and Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81(2):345–381.
- Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Joshua B Tenenbaum. 2023. Evaluating statistical language models as pragmatic reasoners. *arXiv preprint arXiv:2305.01020*.
- Rashid Nizamani, Sebastian Schuster, and Vera Demberg. 2024. Siga: A naturalistic nli dataset of english scalar implicatures with gradable adjectives. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14784–14795.
- Steven T Piantadosi. 2023. Modern language models refute chomsky’s approach to language. *From field-work to linguistic theory: A tribute to Dan Everett*, pages 353–414.
- Zhuang Qiu, Xufeng Duan, and Zhenguang Garry Cai. 2023. Pragmatic implicature processing in chatgpt.
- Chao Sun, Ye Tian, and Richard Breheny. 2024. A corpus-based examination of scalar diversity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 50(5):808.
- Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024. Do large language models understand conversational implicature—a case study with a chinese sitcom. In *China National Conference on Chinese Computational Linguistics*, pages 402–418. Springer.

A Data Access

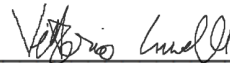
Project work and data are accessible in the [Github repository](#) for this project.

Eigenständigkeitserklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, habe ich in jedem einzelnen Fall unter Angabe der Quelle kenntlich gemacht. *Dies gilt auch für verwendete Zeichnungen, Skizzen, Ton- und Videoaufnahmen sowie graphische Darstellungen.* Ich erkläre mich damit einverstanden, dass meine Arbeit im Verdachtsfall mithilfe einer Plagiatsoftware überprüft wird.

Düsseldorf, 24.04.2025

Ort, Datum



Unterschrift