Human Language Engineering

# Investigating NLI Models' Performance on Figurative Language

**Anna Putina**

Facultat d'Informàtica de Barcelona,
Universtat Politècnica de Catalunya

January, 2024

# Abstract

Natural Language Inference (NLI) plays a pivotal role in training and evaluating models for language understanding. Exploring the proficiency of NLI models in handling inferences related to figurative language, such as idioms and metaphors, adds an intriguing dimension, given their widespread usage in language. This project[1] utilizes the IMPLI dataset[2], which consists of paired English sentences featuring idioms and metaphors.

The evaluation specifically targets NLI models based on BERT and RoBERTa, fine-tuned on the widely adopted MNLI dataset. A comparative analysis of these models demonstrates their reliable ability to detect entailment relationships between figurative phrases and their literal counterparts. However, a noteworthy limitation surfaces as these models exhibit suboptimal performance on similarly structured examples intentionally crafted to be non-entailing. This observation highlights the current constraints of NLI models in effectively comprehending figurative language nuances.

# 1 Introduction

## 1.1 Understanding Natural Language Inference

Natural Language Inference, the task of predicting the connection between two text fragments, has long been a cornerstone in the realm of natural language processing (NLP). In this report, I embark on an exploration of the challenges and intricacies that arise when NLP confronts the subtleties of figurative language, with a specific focus on idioms and metaphors.

## 1.2 Figurative Language and NLP

Figurative language, encompassing idioms, metaphors, and more, introduces a layer of complexity to NLI by deviating from the literal compositional meaning. This departure between speaker intent and literal interpretation significantly influences a myriad of NLP tasks, ranging from sentiment analysis to political discourse analysis. As I explore how figurative language affects these tasks, the crucial reasons to examine and improve NLP systems for better handling the subtleties of figurative expressions are revealed.

| | |
|---|---|
| Idioms | Jamie was *pissed off* this afternoon.<br>→ Jamie was *irritated* this afternoon |
| | There's a marina down *in the docks.*<br>+ There's a marina down *under scrutiny.* |
| Metaphors | The *hearts of men were softened.*<br>→ The *men were made kindler and gentler.* |
| | The gun *kicked* into my shoulder.<br>→ The *mule* kicked into my shoulder. |

**Table 1.** Examples of entailment (→) and nonentailment pairs (↛) from the IMPLI dataset.

## 1.3 Challenges in Figurative Language Processing

In the ever-evolving landscape of large-scale pre-training and transformer-based models, there's a noticeable gap in our grasp of how these models deal with figurative and creative language. Additionally, the lack of comprehensive datasets for assessing NLI in the context of figurative language adds complexity. Building datasets that capture the rich interplay between literal and figurative language is a demanding task. In this project, I worked with a newly created dataset tailored for addressing figurative language. Employing NLI as a perspective, my aim is to explore the subtle dynamics between figurative language and machine comprehension. The objective is to gain insights into how current NLP systems navigate the complexities of figurative expressions, uncovering both strengths and limitations.

# 2 Dataset

The IMPLI[3] dataset used in this project is designed for exploring figurative language. It intricately pairs idiomatic and metaphoric sentences with both entailing and non-entailing counterparts, combining silver pairs generated automatically and manually composed gold pairs. Expert annotators, proficient in English and well-versed in figurative language, played a pivotal role in crafting the dataset.

The creation of the dataset involves various techniques, explicitly outlined by the authors:

- Silver Pairs Generation
  - Annotators craft phrase definitions for figurative expressions.
  - Definitions are automatically inserted into relevant contexts, generating a substantial number of entailment and non-entailment pairs.

[1] Implementation details can be found at: `https://github.com/an-eve/nlp-nli-idioms`.

[2] Dataset and all related resources are publicly available at: `https://github.com/UKPLab/acl2022-impli`

- Idiomatic Pairs
  - Utilization of three corpora (MAGPIE, PIE, SemEval) containing sentences with idiomatic expressions (IEs) labeled as figurative or literal.
  - Manual corrections made by annotators to ensure correctness and syntactical compatibility.
  - Idiomatic pairs constructed by replacing definitions into original sentences.
- Adversarial definitions
  - Adversarial definitions integrated into figurative sentences.
  - Pairs created where the premise is an idiom used figuratively, and the hypothesis attempts to rephrase the idiom literally, yielding non-entailments.
- Metaphoric Pairs
  - Minimal metaphoric expressions (MEs) collected.
  - Annotators modify MEs to create literal counterparts.
  - Modified MEs replaced into sentences, resulting in entailing pairs with the metaphoric sentence entailing the literal.
- Manual Creation of Gold Pairs
  - Annotators rewrite figurative sentences literally, collecting gold standard paraphrases for idiomatic and metaphoric contexts.
  - Annotators write non-entailed hypotheses, preserving lexical overlap while removing the main figurative element.
- Antonyms
  - Annotators replace key words in manually elicited definitions with their antonyms.
  - Sentences with antonym replacements negate the original figurative meaning, serving as non-entailment pairs.

I chose this dataset because the multi-faceted approach ensures the IMPLI dataset's richness, diversity, and relevance for evaluating NLI systems in the context of figurative language.

# 3 Experiments

## 3.1 Objectives

Employing the IMPLI dataset, my objective is to investigate two central inquiries through NLI regarding the proficiency of language models in accurately grasping and portraying figurative language.

These pivotal questions encompass:

1. Evaluation of pre-trained models performance on figurative entailments and non-entailments
   - To what degree do pre-trained models demonstrate competence in handling both figurative entailments and non-entailments?
   - This assessment seeks to gauge the effectiveness of pre-trained models in capturing the intricate relationships inherent in figurative language expressions.
2. Analysis of the impact of including idiomatic pairs in training data on model performance
   - How does the introduction of idiomatic pairs into the training data influence the overall performance of language models?
   - This exploration aims to uncover potential enhancements or challenges introduced by the incorporation of idiomatic expressions, shedding light on the adaptability of models to the nuances of idiomatic language.

These questions serve as the focal point of my investigation.

## 3.2 Implemenation

To address the posed inquiries, I followed a systematic approach:

1. Model Selection:

   I opted to explore BERT[1] (Bidirectional Encoder Representations from Transformers) and RoBERTa[2] (Robustly Optimized BERT Approach). This choice was driven by the intrigue surrounding how these models handle figurative language. While both models share a foundation in transformer architecture and bidirectional pre-training, RoBERTa deviates in its training methodology. These nuanced distinctions, including the abandonment of the next sentence prediction task, dynamic masking, and the use of larger mini-batches, are anticipated to influence performance and robustness compared to BERT. By investigating their responses to figurative language, I aim to discern which model might be better suited for this purpose.

2. Fine-tuning on MultiNLI:

   I conducted fine-tuning on both selected models using the MultiNLI[4] dataset, with entailments as the positive class and all others as negatives. Both models achieved accuracy higher than 90% on the validation sets.

3. Evaluation on IMPLI Dataset:

   The subsequent step involved evaluating both models on the IMPLI dataset. The dataset was divided based on entailment or non-entailment and the method used to create pairs (adversarial definition, replacement in a literal context, etc.).

4. Incorporating Idioms into Training:

   To assess the impact of incorporating idioms into training, I split the idiom data by phrase types. A set of idiomatic expressions was kept separate as test data to evaluate whether the model could effectively handle novel, unseen phrases. The goal was to determine whether poor performance was attributed to the absence of these expressions in training or if there were inherent limitations in the models' ability to represent figurative language.

   For each task, the data was divided into 10 folds, with the idiomatic expressions in the test set not presented in the training sets. The models, already fine-tuned on MultiNLI, were incrementally fine-tuned for each fold. The evaluation included assessing the fine-tuned models on the entire test set, as well as the entailment and non-entailment partitions.

# 4 Results and Reasoning

## 4.1 Results

After completing the aforementioned steps, the obtained results can be summarized as follows:

1. Idiomatic Entailments vs. Non-entailments:

   Idiomatic entailments proved relatively manageable to classify, whereas non-entailments posed a more formidable challenge. Silver pairs generated through adversarial definitions exhibited heightened difficulty due to their substantial lexical overlap. On the other hand, replacing idiomatic expressions into literal samples presented a comparatively easier task. The clash between the idiomatic definition and the original premise accentuated, leading to more discernible non-entailment predictions.

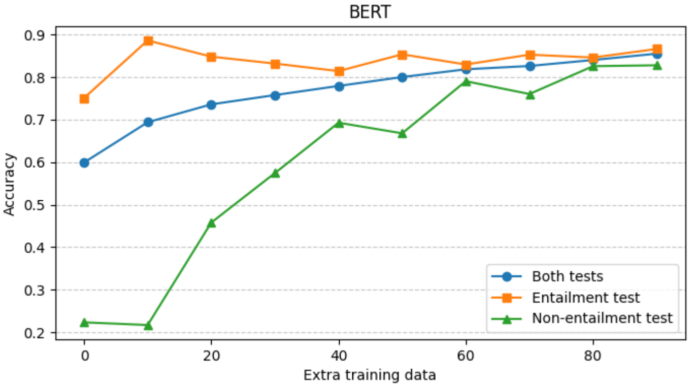

**Figure 1.** Performance of the *bert-base* models as more idiom examples are added to the training data.

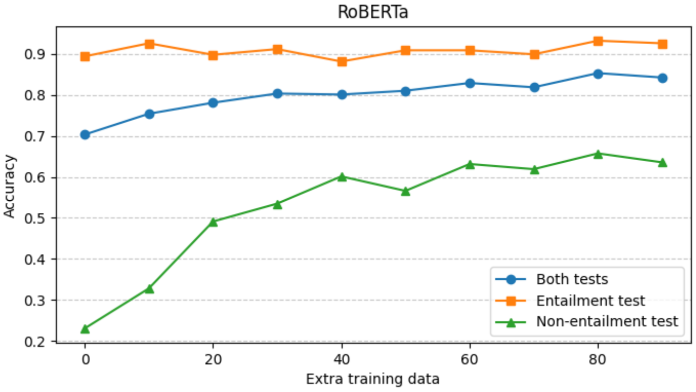

**Figure 2.** Performance of the *roberta-base* models as more idiom examples are added to the training data.

2. Metaphors vs. Idioms:

   Generally, both models found tasks related to metaphors more tractable compared to those associated with idioms.

3. Model Comparison (BERT vs. RoBERTa):

   Surprisingly, RoBERTa outperformed BERT in handling entailment pairs, while BERT demonstrated superior performance with non-entailment pairs. This observation is intriguing, considering RoBERTa's status as an improved version of BERT.

4. Impact of Additional Training Data:

   The inclusion of data into training did not lead to substantial improvements in model performance on figurative language tasks. Notably, a significant enhancement was observed specifically for non-entailment pairs when adding initial portions of data

|  | MNLI |  | IMPLI |  |  | Idioms |  |  |  |  |  | Metaphors |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | m | mm | full | $\to$ | $\nrightarrow$ | $\to$ S | $\nrightarrow$ $S^l$ | $\nrightarrow$ $S^d$ | $\to$ G | $\nrightarrow$ $G^a$ | $\nrightarrow$ G | $\to$ S | $\to$ G | $\nrightarrow$ G |
| bert-base | .901 | .901 | .637 | .727 | .444 | .726 | .679 | .403 | .775 | .653 | .283 | .961 | .822 | .858 |
| roberta-base | .921 | .921 | .704 | .846 | .401 | .844 | .507 | .374 | .894 | .640 | .315 | .972 | .889 | .783 |

**Table 2.** Accuracy on MNLI and IMPLI pairs, divided into silver (S) and gold (G) datasets. $S^l$ Silver non-entailment based on replacement in literal contexts, $S^d$ Silver non-entailment based on adversarial definitions, $G^a$ Gold non-entailment based on antonyms.

into training. However, these results should be interpreted cautiously due to the challenges elaborated in the subsequent section.

5. Distinct Behavior of BERT and RoBERTa with Additional Data:

   Notably, there is a discernible difference in the learning patterns of BERT and RoBERTa when additional data is introduced into training. BERT appears to adeptly handle both entailment and non-entailment pairs with comparable proficiency. In contrast, RoBERTa exhibits a persistent gap in performance between these two categories, even with the augmented training data.

These observations provide valuable insights into the nuanced behavior of BERT and RoBERTa when confronted with figurative language challenges.

### 4.2 Difficulties

1. Fine-tuning large language models is a time-consuming process. I trained both BERT and RoBERTa for three epochs, utilizing the faster GPU option on Google Colab. However, as I progressed to the point of incorporating idiomatic data, I opted to fine-tune the models solely on the IMPLI dataset (approximately 24k pairs), excluding MNLI due to constraints in cost and time. This decision might have implications for the robustness and reliability of the results.

2. The division of the IMNLI dataset into training and testing sets, ensuring that idiomatic expressions (IEs) in the test set were not present in the training set, posed a considerable challenge. While I devised a script to accomplish this task, manual cleaning was necessary. Despite these efforts, some data leaks may persist due to the inherent flexibility of language—expressions can be used in different tenses, divided by adverbs, etc. For future endeavors, including a column specifying the idiomatic expression for each pair could significantly simplify data management.

These difficulties are noteworthy as they could potentially impact the integrity of the results.

## 5 Conclusions

The evaluation of two NLI models' proficiency in handling figurative language was conducted.

The findings revealed that widely used MNLI models excel in managing entailment tasks, and metaphoric expressions pose a relatively manageable challenge. However, dealing with non-entailment idiomatic relationships proved to be more intricate. Moreover, the introduction of idiom-specific training data did not alleviate the observed performance issues. This underscores the inherent limitations of current language models in effectively representing certain figurative phenomena, pointing towards potential areas for future model enhancements.

Additionally, it is noteworthy that different language models may exhibit distinct behaviors when confronted with figurative language. This aspect presents an intriguing avenue for exploration in future research, offering the prospect of gaining deeper insights into how various models navigate and interpret figurative expressions.

## References

[1] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: `1810.04805`. URL: `http://arxiv.org/abs/1810.04805`.

[2] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019). arXiv: `1907.11692`. URL: `http://arxiv.org/abs/1907.11692`.

[3] Kevin Stowe, Prasetya Ajie Utama, and Iryna Gurevytch. “IMPLI: Investigating NLI Models’ Performance on Figurative Language”. In: *Proceedings of the 2022 Conference for the Association of Computational Linguistics*. Association for Computational Linguistics, June 2022. URL: `tbd`.

[4] Adina Williams, Nikita Nangia, and Samuel Bowman. “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 1112–1122. URL: `http://aclweb.org/anthology/N18-1101`.