Literary Metaphor Detection with LLM Fine-Tuning and Few-Shot Learning

Spielberg, Marina
s2maspie@uni-trier.de
Trier Center for Digital Humanities (TCDH), Universität Trier, Deutschland

Introduction

The study of literary metaphors plays an integral part in literature-focused disciplines within the humanities. In the field of Natural Language Processing (NLP), computational metaphor detection (MD) has produced a wealth of approaches focusing on everyday metaphors (Ptiček and Dobša 2023). Computational literary MD, however, has received considerably less attention. In their Graph Project, Kesarwani et al. (2017) have applied rule-based and statistical machine learning approaches to an English poetry corpus. The aim of this paper is to take the field of literary MD one step further by using the NLP state-of-the-art approach of fine-tuning Large Language Models (LLMs) on the Graph project's datasets.[1] Considering the small size of the relevant datasets, the few-shot learning approach SetFit is used alongside the more established fine-tuning Transformers approach to compare their performance.

This paper tests two assumptions:

1. Fine-tuning the LLM DistilBERT with the Transformers approach and the LLM all-MiniLM-L6-v2 with the SetFit approach on the Graph project's datasets will yield better results than using the combined statistical and rule-based approach from Kesarwani et al. (2017).
2. The SetFit approach will outperform the Transformers approach.

The paper starts by setting out the theoretical background of MD and then explains the data, method and experimental setup used. It closes with a description of the evaluation results and a discussion.

Background and previous work

Metaphor Detection in NLP

Since the current theories and methods regarding MD come from NLP, it is beneficial to understand prevalent research in this area before focusing on the state of the field in the Digital Humanities. In most of the MD research in NLP, metaphors are understood as a mapping of a source domain like WAR to a target domain like ARGUMENT that can result in the metaphorical linguistic expression "Your claims are **indefensible**" (Lakoff and Johnson 1980, 4). This Conceptual Metaphor Theory is used to detect conversational and novel metaphors (Ptiček and Dobša 2023).

Since 1975, research on MD is ongoing because "the task is not considered solved" (Dankin, Bar, and Dershowitz 2022, 125). The methodologies applied to MD have evolved with the development of computational capabilities starting with rule-based, statistical and machine learning methods (Ptiček and Dobša 2023). The current state-of-the-art method for MD involves utilizing LLMs, often derivations of the BERT model (Devlin et al. 2019), which are based on the Transformer architecture by Vaswani et al. (2017) that allows to fine-tune an existing language model on a specific downstream task like MD (Babieno et al. 2022; Li et al. 2023; Z. Song et al. 2024). Recently, prompt engineering started to be utilized for MD. Instead of labelled data this method uses task-specific formulated prompts (Jia and Li 2024). Chen et al. (2024) expanded the task to include metaphor reason in addition to detection since they found that methods focusing solely on a metaphorical versus literal distinction did not generalise well. Another recent development is to detect metaphor in multi-modal settings, such as memes, where the classification task includes both texts and images (Xu et al. 2024).

Metaphor Detection in the Digital Humanities

In the Digital Humanities, MD focusing on literary texts is understudied. To my knowledge, there are only a handful of papers that concern themselves specifically with this task. Reinig and Rehbein (2019) proposed a supervised machine learning method for MD in German expressionist poetry, while Schneider et al. (2022) developed an unsupervised approach for Middle High German. Toker et al. (2024) used LLMs on their own Early Medieval Hebrew poetry dataset.

In their Graph Project, Kesarwani et al. (2017) developed models to detect poetic metaphors in English, which has neither been done in NLP nor in the Digital Humanities. Diverging from the Conceptual Metaphor Theory, the authors based their metaphor definition on observations by Neuman et al. (2013), who found three metaphor types that signify their metaphoricity by different part-of-speech (POS) sequences. They focused on detecting Type I metaphors, which are comprised of a "Noun-Verb-Noun" sequence and added the sequence "Noun-Verb-Det-Noun" to this concept. An example for Type I metaphor is the sentence "As if the **world were a taxi**" (Kesarwani et al. 2017, 2).

The authors trained and tested on four datasets: Their own PoFo (Poetry Foundation) dataset, comprising 680 sentences that include Type I metaphor scraped from the Poetry Foundation website, the benchmark datasets TroFi by Birke and Sarkar (2006) (6,435 sentences from the Wall Street Journal Corpus) and MOH by Mohammad, Shutova,

and Turney (2016) (647 sentences from WordNet). Finally, they created a fourth concatenated dataset which combines PoFo, TroFi and MOH. See table 1 for an overview of sample sentences labelled "metaphorical" from each dataset.

The methods of the Graph Project mirror the progression of methods in the NLP tradition as the authors experimented with rule-based and statistical approaches (Kesarwani et al. 2017). The F1 scores for their rule-based and machine learning models were 0.669 for PoFo, 0.827 for TroFi, 0.779 for MOH and 0.781 for the concatenated dataset. Tanasescu, Kesarwani, and Inkpen (2018) used deep learning with convolutional neural networks on the concatenated dataset and reached an F1 score of 0.833. Since fine-tuning LLMs proved to be a very successful approach in recent NLP research (e.g., a F1 score of 0.944 on TroFi by Ma et al. 2021), this paper tests whether this method will improve the MD performance on the Graph project's datasets.

Table 1. Dataset domains and sample sentence labeled "metaphorical" from PoFo, TroFi and MOH.

dataset	domain	sample sentence
PoFo	Poetry Foundation - poetry corpus	love is a fiction I must use ,
TroFi	Wall Street Journal - newspaper corpus	And most of the jobs will go to candidates sought out by search firms, not those knocking on headhunters doors
MOH	WordNet - lexical corpus	Her husband often abuses alcohol .

Data, Methodology and Experiments

Based on the assumption that the metaphoricity of a word stems from the context of the whole sentence rather than a single aspect word or POS sequence, I define MD as a sentence-level classification problem, where each sentence within the four datasets is labelled as either "metaphorical" or "literal" (Ma et al. 2021). This approach differs from Kesarwani et al. (2017), who annotate metaphor on a token level. For preprocessing, I reused the cleaned versions of the TroFi and MOH datasets by Su et al. (2020) due to a lack of clear preprocessing information from the Graph project and used the original version of PoFo by Kesarwani et al. In the concatenated dataset there are more literal than metaphorical sentences since it consists of 75% of sentences belonging to TroFi, which suffers from label imbalance, having about 12% more literal than metaphorical examples (figure 1).

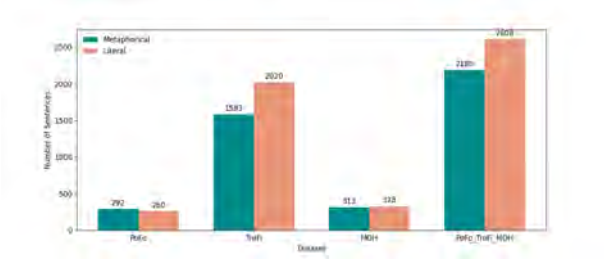

Figure 1. Sentence distribution from the PoFo, TroFi, MOH and concatenated datasets sorted by the labels "metaphorical" and "literal".

The first method to improve the MD performance is fine-tuning the four datasets on the Transformer-based pre-trained LLM DistilBERT (Sanh et al. 2019), which is an efficient variant of BERT. This model is smaller and faster while maintaining 97% of BERTs performance. The training for this Transformers approach is threefold: The model-specific tokenizer maps the dataset's text tokens to indexes, the transformer converts these indexes to contextual embeddings and the pre-trained head is fine-tuned on the MD task (Wolf et al. 2020). The fine-tuning procedure consisted of training with 70% of the data and evaluating with the remaining 30% for each dataset. For hyperparameters, the batch size of 32, the learning rate of 2e-5 and 5 training-epochs yielded best results.

Since Transformer LLMs require fine-tuning on relatively large datasets which is a challenge for literary metaphor datasets, as can be seen from the sizes of the Graph Project's datasets, this paper also employs the SetFit framework (Tunstall et al. 2022). It is designed for few-shot learning, that is, learning with a small number of labelled examples. SetFit operates in two steps: It generates sentence pairs, thereby enlarging the dataset significantly, and then fine-tunes embeddings of these sentences to create a sentence transformer embedding model. Then the dataset is used to train a logistic regression classifier using the fine-tuned embeddings to predict the right labels. The impact of the Sentence Transformer on the dataset size is immense: From 441 samples of the PoFo train dataset, SetFit generated 98366 unique pairs. The all-MiniLM-L6-v2 Sentence Transformer model is used within this framework because it is 5 times faster than larger models while maintaining comparable performance (Reimers et al. 2019). For implementation, the same test-train split ratio, seed and hyperparameters are used as for the Transformers implementation to maintain comparability.

Results

Figure 2 presents the evaluation results of fine-tuning with the Transformers and SetFit methods on the MD task at sentence-level. All reported results are from evaluating the fine-tuned models on the test splits. At least one of the LLM fine-tuning approaches proposed in this paper outperformed the baseline on all datasets except for TroFi. For PoFo

there was a significant performance increase of 12.41% with SetFit (F1 0.752) and 2.84% with Transformers (F1 0.688). The MOH dataset displayed an F1 score of 0.785 with the Transformers approach (0.77% increase). The SetFit approach demonstrated a significant improvement with an F1 score of 0.862 (10.37% increase) on the concatenated dataset. However, the TroFi dataset performed much better with the rule-based and statistical method, surpassing SetFit by 22.25%. As for the comparison of Transformers with SetFit, SetFit achieved better performance on PoFo, TroFi and the concatenated dataset, while Transformers gave better results on the MOH dataset. Overall, no method excelled across all datasets, although the LLM approaches generally performed better.

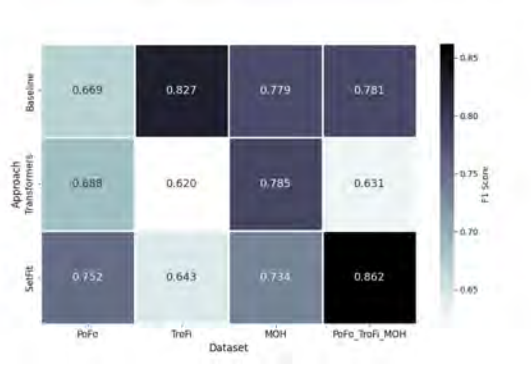

Figure 2. Baseline F1 scores for the PoFo, TroFi, MOH and concatenated datasets as reported by Kesarwani et al. (2017) compared to the F1 scores achieved in this paper during fine-tuning with Transformers and SetFit.

Discussion

The promising results indicate the potential of fine-tuning LLMs for literary MD in general and using the SetFit method for small datasets in particular. The concatenated dataset's heterogeneity and diversity in sentence length and number, complexity, and metaphor domains provided broader contextual information, explaining its high F1 score. Despite these positive outcomes fine-tuning did not achieve better performance on TroFi.[2]

TroFi's suboptimal performance could stem from training difficulty caused by the sentence-level approach and label imbalance. However, these arguments can be refuted by TroFi's substantial representation in the concatenated dataset (75.1%, table 1), which showed exceptionally good results. Thus, TroFi's characteristics must have significantly contributed to these positive results.

The chosen LLM models might also contribute to the low result. Being condensed models, DistilBERT and all-MiniLM-L6-v6 are useful for initial experiments but might not generalise well on a domain-specific task like MD. Using SetFit on a state-of-the-art Sentence Transformer model like all-mpnet-base-v2 (Song et al. 2020) may help improve results.

However, the core difficulty of finding concrete explanations for the poor TroFi results and the excellent concatenated dataset and PoFo results is the interpretability limitation of LLMs due to their "black-box" nature. Contrary to statistical machine learning approaches, no information is provided about which features contribute most to the classification task during LLM fine-tuning, making the reasoning behind the model's classification result not entirely comprehensible (Dobson 2023, 431). This is especially concerning for the Digital Humanities, where understanding the domain is just as important as raw performance. Ablation techniques and visualisations of attention weights could help understanding how model output was created.

Conclusion

This paper contributed to the understudied task of literary MD by applying state-of-the-art NLP methodology, like fine-tuning the Transformer model DistilBERT and few-shot learning with the Sentence Transformer approach, on four literary metaphor datasets. Metaphor was defined quite narrowly as consisting of one of two specific POS sequences. The evaluation baseline was the combined rule-based and statistical approach of Kesarwani et al. (2017).

The results demonstrate performance increases in F1 score for the fine-tuning approach over the baseline and even more so for the SetFit methods, especially for PoFo (12.41% increase) and the concatenated dataset (10.37% increase). However, improvement was not observed for the TroFi dataset, which could stem from sentence complexity and label imbalance or from small model sizes.

These findings emphasise that while the current practise of fine-tuning LLMs for linguistic MD can also yield good results for literary MD and that SetFit is a valuable tool for small datasets, these methods do not guarantee improved performance. Due to the black-box nature of LLMs they might not be the right tool for literary scholars, who prioritise interpretability.

Further work needs to be performed to establish whether larger models could optimise the work done in this paper. Future studies on literary MD could focus on creating larger datasets with different kinds of metaphors or employ prompt engineering. The fine-tuned models could be deployed to build interactive tools for teaching and studying metaphors in educational settings.

Fußnoten

1. The code, datasets and outputs for this paper are publicly available on GitHub: https://github.com/ma-spie/LLM_metaphor_detection.
2. While judging result comparability it should be noted that this paper introduced variations in dataset setup compared to the approach by Kesarwani et al. (2017).

Bibliographie

Babieno, Mateusz, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. "Miss Roberta Wilde: Metaphor Identification Using Masked Language Model with Wiktionary Lexical Definitions." *Applied Sciences* 12 (2081). https://doi.org/10.3390/app12042081.

Birke, Julia, and Anoop Sarkar. 2006. "A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language." *11th Conference of the European chapter of the association for computational linguistics*, 329–36. https://aclanthology.org/E06-1042.

Chen, Puli, Cheng Yang, and Qingbao Huang. "Merely Judging Metaphor Is Not Enough: Research on Reasonable Metaphor Detection." In *Findings of the Association for Computational Linguistics: EMNLP 2024*, edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 5850–60. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.336.

Dankin, Lena, Kfir Bar, and Nachum Dershowitz. 2022. "Can Yes-No Question-Answering Models Be Useful for Few-Shot Metaphor Detection?" In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, edited by Debanjan Ghosh, Beata Beigman Klebanov, Smaranda Muresan, Anna Feldman, Soujanya Poria, and Tuhin Chakrabarty, 125–30. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.flp-1.17.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the NAACL-HLT 2019*, 4171–86. http://arxiv.org/pdf/1810.04805.

Dobson, James E. 2023. "On Reading and Interpreting Black Box Deep Neural Networks." *International Journal of Digital Humanities* 5 (2): 431–49. https://doi.org/10.1007/s42803-023-00075-w.

Jia, Kaidi, and Rongsheng Li. 2024. "Enhancing Metaphor Detection Through Soft Labels and Target Word Prediction." *31st Conference on Neural Information Processing Systems*. http://arxiv.org/pdf/2403.18253.

Kesarwani, Vaibhav, Diana Inkpen, Stan Szpakowicz, and Chris Tanasescu. 2017. "Metaphor Detection in a Poetry Corpus." In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, edited by Beatrice Alex, Stefania Degaetano-Ortlieb, Anna Feldman, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, 1–9. Stroudsburg, PA, USA: Association for Computational Linguistics. https://aclanthology.org/volumes/W17-22.

Lakoff, George, and Mark Johnson. *Metaphors we live by*. 1980. Chicago: University of Chicago Press.

Li, Yucheng, Shun Wang, Chenghua Lin, Frank Guerin, and Loic Barrault. 2023. "FrameBERT: Conceptual Metaphor Detection with Frame Embedding Learning."In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 1558–63. https://doi.org/10.18653/v1/2023.eacl-main.114.

Ma, Weicheng, Ruibo Liu, Lili Wang, and Soroush Vosoughi. 2021. "Improvements and Extensions on Metaphor Detection." In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, edited by Michael Roth, Reut Tsarfaty, and Yoav Goldberg, 33–42. Stroudsburg, PA, USA: Association for Computational Linguistics. https://aclanthology.org/2021.unimplicit-1.5.

Mohammad, Saif, Ekaterina Shutova, and Peter Turney. 2016. "Metaphor as a Medium for Emotion: An Empirical Study." In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 23–33. https://doi.org/10.18653/v1/S16-2003.

Neuman, Yair, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, and Ophir Frieder. 2013. "Metaphor Identification in Large Texts Corpora." *PloS one* 8 (4): 1-9. https://doi.org/10.1371/journal.pone.0062343.

Ptiček, Martina, and Jasminka Dobša. 2023. "Methods of Annotating and Identifying Metaphors in the Field of Natural Language Processing." *Future Internet* 15 (6): 1-28. https://doi.org/10.3390/fi15060201.

Reimers, Nils, Omar Espejel, Pedro Cuenca, and Tom Aarsen. n.d. "Pretrained Models." *Sentence-Transformers Documentation*. Accessed July 23, 2024. https://sbert.net/docs/sentence_transformer/pretrained_models.html.

Reinig, Ines, and Ines Rehbein. "Metaphor Detection for German Poetry." In *Preliminary Proceedings of the 15th Conference on Natural Language Processing*, 149–60. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-93163.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter.", 1–5. https://doi.org/10.48550/arXiv.1910.01108.

Schneider, Felix, Sven Sickert, Phillip Brandes, Sophie Marshall, and Joachim Denzler. 2022. "Metaphor Detection for Low Resource Languages: From Zero-Shot to Few-Shot Learning in Middle High German." In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, edited by Archna Bhatia, Paul Cook, Shiva Taslimipoor, Marcos Garcia, and Carlos Ramisch, 75–80. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.mwe-1.11.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. "MPNet: Masked and Permuted Pre-Training for Language Understanding." In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Vancouver, BC, Canada: Curran Associates Inc. 16857–67. https://doi.org/10.48550/arXiv.2004.09297.

Song, Ziqi, Shengwei Tian, Long Yu, Xiaoyu Zhang, and Jing Liu. 2024. "Multi-Task Metaphor Detection

Based on Linguistic Theory." *Multimed Tools Appl*. https://doi.org/10.1007/s11042-023-18063-1.

Su, Chuandong, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. "DeepMet: A Reading Comprehension Paradigm for Token-Level Metaphor Detection." In *Proceedings of the Second Workshop on Figurative Language Processing*, edited by Beata B. Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee Wee, Anna Feldman, and Debanjan Ghosh, 30–39. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.figlang-1.4.

Tanasescu, Chris, Vaibhav Kesarwani, and Diana Inkpen. 2018. "Metaphor Detection by Deep Learning and the Place of Poetic Metaphor in Digital Humanities." *The thirty-first international flairs conference*, 122–27. https://aaai.org/papers/122-flairs-2018-17704.

Toker, Michael, Oren Mishali, Ophir Münz-Manor, Benny Kimelfeld, and Yonatan Belinkov. 2024. "A Dataset for Metaphor Detection in Early Medieval Hebrew Poetry." https://aclanthology.org/2024.eacl-short.39.

Tunstall, Lewis, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. "Efficient Few-Shot Learning Without Prompts." *36th Conference on Neural Information Processing Systems*, 1–14. https://doi.org/10.48550/ARXIV.2209.11055.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30:1–11. https://doi.org/10.48550/arXiv.1706.03762.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac. 2020. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*: *System Demonstrations*, edited by Qun Liu and David Schlangen, 38–45. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Xu, Yanzhi, Yueying Hua, Shichen Li, and Zhongqing Wang. "Exploring Chain-of-Thought for Multi-Modal Metaphor Detection." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Lun-Wei Ku, Andre Martins, and Vivek Srikumar, 91–101. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.6.