Metaphor and Figurative Language Detection Using LLMs

Preda Alexandru-Florin

Abstract

This project investigates the application of state-of-the-art large language models (LLMs) for metaphor and figurative language detection. I explored the **DeepSeek-R1** with 14B parameters model to process and analyze sentences from the **VUA metaphor detection** dataset. Using a proof-of-concept implementation, I evaluate the model's ability to classify target words as metaphorical or literal. My analysis discusses the methodology and outlines directions for future improvements.

Analysis of the Main Idea

Metaphor detection is about finding words that are used in a creative or abstract way, instead of their usual, literal meaning. It's a tricky task in natural language processing. Thanks to advanced language models that are good at understanding context, this challenge can be tackled more effectively. This project tests how well a new model, **DeepSeek-R1:14b**, can recognize metaphors by using its ability to understand context. I picked **DeepSeek-R1:14b** because it's cutting-edge and a good choice for testing on this kind of detailed language task. To help the model, I give it structured prompts and annotated examples to guide its predictions.

Related Work: State of the Art

In the past, metaphor detection relied on rule-based systems that used manually crafted language rules. As machine learning advanced, researchers began using models trained with custom linguistic features and word embeddings. The introduction of transformer-based models like BERT and GPT significantly improved metaphor detection by providing better contextual understanding. Today, the most effective approaches often involve fine-tuned versions of these transformers or hybrid methods that combine deep learning with symbolic reasoning. **DeepSeek-R1:14b** belongs to this new generation of advanced language models, making it a strong candidate for metaphor detection due to its ability to process nuanced language and context effectively.

One notable recent approach is **ContrastWSD**, introduced by Elzohbi and Zhao. This model enhances metaphor detection by integrating **Word Sense Disambiguation (WSD)** and the **Metaphor Identification Procedure (MIP)**. Unlike standard transformer-based models that rely entirely on contextual embeddings, ContrastWSD explicitly compares a word's contextual meaning with its basic meaning to determine if it is being used metaphorically. By using WSD-derived word senses, this approach improves detection accuracy compared to models that only incorporate basic definitions or external knowledge. Evaluations on benchmark datasets show that ContrastWSD outperforms other methods, highlighting the benefits of incorporating structured linguistic insights into metaphor detection.

This research highlights the potential of combining large-scale language models with explicit sense disambiguation techniques. Inspired by these advancements, my study explores the capabilities of DeepSeek-R1:14b in metaphor detection, assessing whether its strong contextual comprehension can match or exceed existing state-of-the-art methods.

Methodology

For this project, I used the metaphor-detection-vua-wsd-augmented dataset. It contains sentences with labels for things like part of speech, word meanings, and definitions. Since the DeepSeek-R1:14b model requires a lot of computing power, I worked with a smaller sample of about 100 randomly chosen rows to keep it manageable and ensure a mix of data.

The DeepSeek-R1:14b model was accessed using the Ollama server, which allowed interaction through API calls. To test the model, prompts were carefully structured to give clear and detailed input. Each prompt included the sentence being analyzed, the target word, it's part of speech, word sense, word index, and a definition. The model was then asked if the target word was used metaphorically, and it responded with a simple yes or no.

The API interaction was handled by a function called **get_response_content**, which sent a **POST** request to the server at **http://localhost:11434/api/chat**. It passed the model's name, a system role message, and the user's prompt. If the server returned a valid response, the function extracted and returned the relevant content. Prompts were generated using the **create_prompt** function, which formatted the required details for each word into a structured question for the model to evaluate.

To evaluate the model's performance, a confusion matrix was computed by comparing its predictions to the ground truth annotations in the dataset. This allowed for a detailed analysis of how well the model identified metaphors. Based on the confusion matrix, key performance metrics such as precision, recall, and F1-score were then calculated to assess the model's effectiveness.

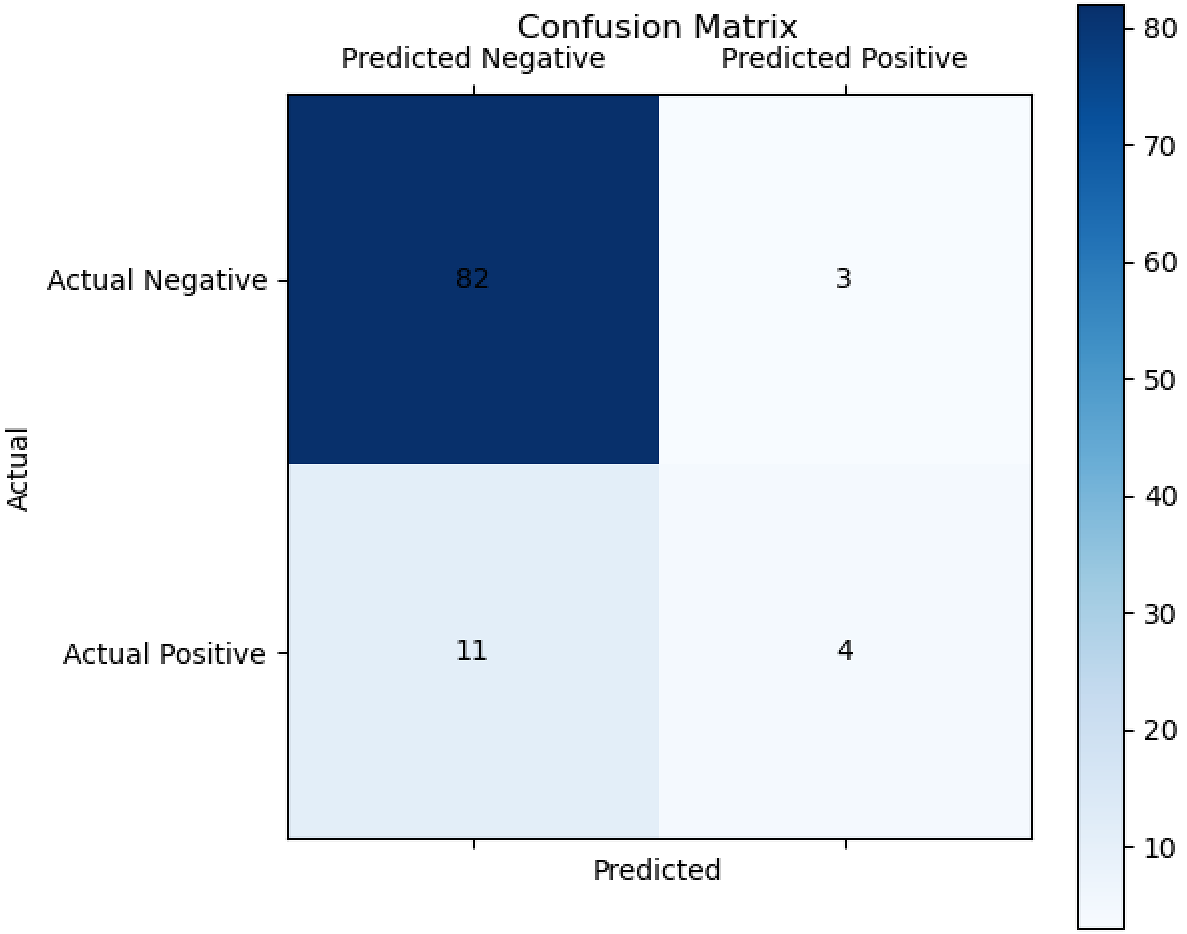
Experimental Results

Model	Rec	Prec	F1
SOTA Results			
MelBert	77.5	79.87	78.66
MsW_cos	77.88	80.31	79.07
FrameBERT	76.78	79.33	78.03
ContrastWSD	78.85	80.16	79.50
MY RESULTS			
Llama3.2-1b	73	18	29
Llama3.2-3b	07	50	12
deepseek-r1:8b	13	22	17
Deepseek-r1:14b	26.67	57.14	36.36

The experimental results show that the DeepSeek-R1:14b model, with a recall of 26.67, precision of 57.14, and F1-score of 36.36, outperformed smaller models like DeepSeek-R1:8b and other configurations such as Llama3.2. Despite falling short of state-of-the-art models like MelBERT and ContrastWSD, the results highlight the potential of DeepSeek-R1 in handling nuanced tasks like metaphor detection.

A key observation is that larger models like DeepSeek-R1:14b performed better, which suggests that scaling up model parameters improves performance. This is promising given that DeepSeek-R1 has models with parameters up to 671 billion. Future experiments with these larger models could significantly improve results, as seen from the trend in parameter scaling.

Additionally, the methodology used here is unique. Instead of using DeepSeek-R1 to extract features for training a classification head, such as a multi-layer perceptron (MLP) or a traditional model like SVM, the model was directly prompted for predictions. While this unconventional approach simplifies the pipeline, it may limit the classification accuracy compared to using a custom head for task-specific tuning. However, due to the unavailability of DeepSeek-R1's internals for feature extraction, this method was necessary. Future work could explore traditional classification approaches with DeepSeek-R1’s features once the model becomes more accessible.



DeepSeek-R1:14B Confusion Matrix

Conclusion

This study explored the capabilities of the DeepSeek-R1:14b model for metaphor and figurative language detection, utilizing a proof-of-concept implementation. While the results fell short of state-of-the-art models such as MelBERT and ContrastWSD, the performance of DeepSeek-R1:14b—achieving a precision of 57.14 and an F1-score of 36.36—highlights its potential for nuanced language tasks. The findings also demonstrate that increasing model size improves performance, as seen in the comparison between DeepSeek-R1:8b and DeepSeek-R1:14b. This trend suggests that exploring even larger models in the DeepSeek-R1 series, such as those with up to 671 billion parameters, could yield significant advancements in metaphor detection.

The study's approach was unconventional, using the model as a prompt-based classifier instead of extracting features for a separate classification head, such as an MLP or SVM. This method simplified the pipeline but may have limited classification accuracy. Future research could address this by integrating traditional classification techniques, leveraging the model's features once they become more accessible.

Overall, this work underscores the promise of large language models in handling complex linguistic tasks and paves the way for further exploration of scalable and hybrid approaches to improve metaphor detection.

Bibliography:

- Mohamad Elzohbi and Richard Zhao. 2024. ContrastWSD: Enhancing Metaphor Detection with Word Sense Disambiguation Following the Metaphor Identification Procedure. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 3907–3915, Torino, Italia. ELRA and ICCL.