DRMD: Explainable Depression Detection Based on Metaphorical Conceptual Mapping

Anonymous ACL submission

Abstract

Depression is a pervasive mental health issue affecting millions globally, and social media has become a key platform for individuals to express their emotional struggles. The textual content shared by individuals with depression contains valuable insights into their mental states, yet analyzing such data presents challenges due to the complexity of indirect expressions, including metaphors. These metaphorical expressions can provide crucial insights into the psychological states of individuals with depression and play an important role in therapeutic contexts. This paper addresses the challenge of detecting depression by leveraging metaphorical information. We introduce a novel, publicly available Depression-related metaphor dataset (DRMD), which contains social media posts from individuals with depression, along with metaphorical labels and conceptual source domain mappings. This dataset is used to fine-tune large language models (LLMs), integrating metaphorical features to enhance the models' depression detection performance. Our results demonstrate that the fine-tuned models with metaphorical information not only improve detection accuracy but also generate high-quality explanations for detection outcomes, utilizing metaphorical expressions to offer deeper insights into the mental states of individuals. This work highlights the potential of metaphorical analysis in mental health diagnostics and provides a foundation for future research in automated depression detection and explanation generation. The dataset is publicly available[1].

1 Introduction

Depression is a serious health and social issue that currently affects the physical and mental health of over 350 million people. Identifying and diagnosing its early symptoms has become a crucial public-health topic. With the development of network technology, social media platforms have gradually become important data sources for mental health research. Depressed patients often use social media to express their emotional experiences, and the text data they post provides a unique window for the analysis of mental illnesses (Guntuku et al., 2017;Benton et al., 2017). Although existing research has made progress in automatic depression detection (Suhara et al., 2017;Tadesse et al., 2019;Zogan et al., 2021;Gui et al., 2019;Lin et al., 2020;Ji et al., 2021), it generally neglects the key language feature of metaphorical expressions. Clinical studies show that patients with mental illnesses tend to use indirect means of expression like metaphors to convey emotions (Magaa, 2019). This language phenomenon is particularly prominent among the depressed population (Coll-Florit et al., 2021;Roystonn et al., 2021). Research has proven that analyzing metaphorical concept mappings can not only reveal patients' subconscious mental states but also promote the effectiveness of therapeutic communication (Kopp, 2013;Siegelman, 1993). This offers important inspiration for enhancing the interpretability of automatic detection models.

With the rapid development of large language model (LLM) technology, it has shown potential in various downstream tasks (Cao et al., 2023;Ji et al., 2023;Fei et al., 2023;Zhang et al., 2023), including mental health analysis (Brown et al., 2020;Wei et al., 2022;Yang et al., 2023) and metaphor detection (Yang et al., 2024b). Therefore, we integrate metaphorical features into the reasoning process of LLMs, enabling them to capture the deep-seated psychological information contained in metaphorical expressions. Currently, fine-tuning is the most effective method to improve the performance of LLMs in specific tasks. However, this approach faces a key issue. Fine-tuning requires high-quality supervised training data, yet

[1] https://anonymous.4open.science/r/DRMD-AC6B

existing public datasets for depression detection tasks generally lack metaphor annotations (Pirina and Çöltekin, 2018;infamouscoder, 2022;Cohan et al., 2018).

The main contributions of this work are summarized as follows:

1. We constructed a new publicly available dataset consisting of textual content posted by individuals with depression on social media platforms. The dataset includes the metaphors used and the conceptual source domains of their metaphorical mappings.

2. We enhanced the performance of LLMs in depression detection by introducing metaphorical features and generated high-quality explanations for detection results based on conceptual source domain mappings.

3. We constructed a diversified instruction set based on the dataset to fine-tune two baseline models. The results show that the fine-tuned models demonstrated improved depression detection capabilities and significantly enhanced performance in generating high-quality explanations.

2 Related work

2.1 Depression Detection Dataset

Datasets play an important role in mental health analysis tasks. The differences in existing mental health datasets mainly lie in the related tasks, data sources, sample sizes, and annotation content.In terms of tasks, most datasets are mainly applied to classification or detection tasks such as (Naseem et al., 2022;Pirina and Çöltekin, 2018;Shen et al., 2017;Turcan and McKeown, 2019;Coppersmith et al., 2015;Pérez et al., 2023), while a small number focus on simulating psychological counseling (Sun et al., 2021;Lahnala et al., 2021).In terms of data sources, existing datasets mainly come from social media platforms such as Reddit, Twitter, and Weibo. The second source is from mental health counseling or simulated clinical interview dialogues from real-world settings (Liu et al., 2023;Sun et al., 2021;Yao et al., 2022), but samples collected through this approach are generally smaller. For example, D4 (Yao et al., 2022) contains only 1,339 expert-reviewed conversations, whereas data collected from Twitter for the MDDL (Shen et al., 2017) includes 292,564 tweets from 1,402 depressed users.In terms of annotation content, most datasets' annotations mainly consist of binary classification labels for depressive/non-depressive texts (Turcan and McKeown, 2019;Pirina and Çöltekin, 2018), or they categorize different text data by labeling depressed/non-depressed users (Shen et al., 2017). Furthermore, in (Naseem et al., 2022), the severity of depression in depressed users was classified .

Clearly, the annotation content in existing publicly available datasets is insufficient to provide theoretical support for LLMs to generate persuasive explanations for their detection results. Therefore, the metaphors and their mapped conceptual source domains annotated in our dataset are crucial for enhancing the explanatory capabilities of LLMs .

2.2 Depression metaphor

In Conceptual Metaphor Theory (CMT) (Lakoff and Johnson, 1980), people can more easily understand complex and abstract concepts by mapping relatively abstract or unexperienceable concepts (target domains) to familiar and relatively concrete concepts (source domains).

Based on this theory, (Barcelona Sánchez et al., 1986) conducted a study on depression metaphors in English and first summarized several metaphorical concepts for depression, such as mapping depression to source domains like "BURDEN","LIVING ORGANISM","ENEMY"and "BOUNDED SPACE" to depict personal feelings. Additionally, an analysis of the language used by individuals with depression revealed that the majority of the metaphors they used mapped depression to the source domain "DEPRESSION IS DESCENT", with smaller portions using "DEPRESSION IS DARKNESS", "DEPRESSION IS WEIGHT" and "DEPRESSION IS CAPTOR"(McMullen and Conway, 2002). Furthermore, (Charteris-Black, 2012) introduced the concept of containment and constraint as metaphorical source domains for the first time. Common expressions of this domain include depicting the individual with depression as a container for negative emotions or likening depression to a container that surrounds and constrains the individual.

And by analyzing the content posted by users on a large number of social media platforms(Tonon, 2020), they identified 25 metaphorical source domains associated with depression. Among the most prominent were "DEPRESSION IS A DOWNWARD MOVEMENT","DEPRESSION

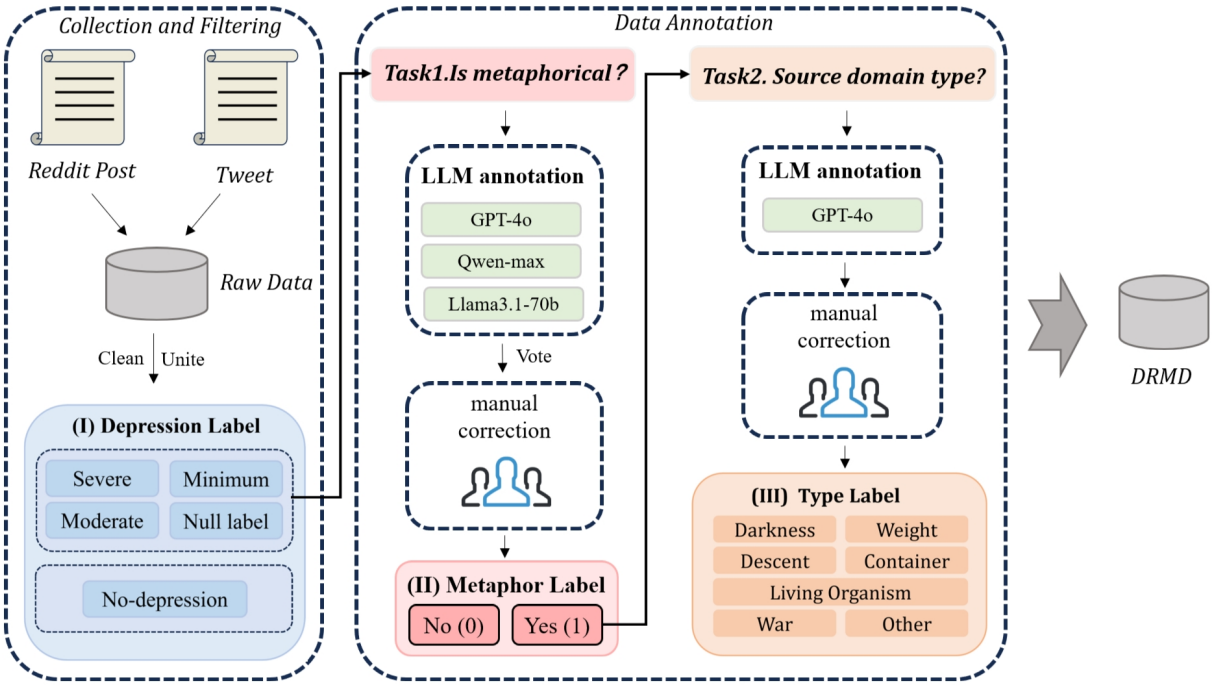

Figure 1: DRMD dataset construction process.

IS DARKNESS","DEPRESSION IS CONFLICT","DEPRESSION IS A BOUNDED SPACE" and "THE DEPRESSED PERSON IS A CONTAINER".

In this paper, we identify the limitations of existing work in depression detection. Previous depression detection studies have mainly focused on depression detection or related classification tasks. However, due to the lack of domain knowledge training, the field of law has been less involved in psychological health analysis tasks. Therefore, we combine previous research on depression metaphors and use metaphorical information to interpret detection results.

3 Dataset

In this section, we introduce the process of constructing the dataset. We integrated four publicly available datasets for depression detection and used large language models to further annotate the metaphorical information in these datasets. After conducting a manual evaluation of the annotation results, we obtained the final dataset. The dataset construction process is illustrated in Fig.1.

3.1 Data Collection and Filtering

This dataset aims to help explore the relationship between the metaphors used in the content posted by users with depressive tendencies on social media and their mental health status. We began by collecting publicly available datasets for depression detection and, based on the data quality and labeling, selected four datasets to form the original dataset. The data in these four datasets were collected from the Reddit platform. The final original dataset consists of a total of 55,173 depression-related data points and 7,482 non-depression-related data points. The labels and sizes of these four datasets are shown in Tab.1.

Dataset	Label	Number
Naseem et al., 2022	Binary Depression	41819
Sampath and Durairaj, 2022	Binary Depression / Depression degree	10352
Pirina and Çöltekin, 2018	Binary Depression	3009
infamouscoder, 2022	Binary Depression	7729

Table 1: The original datasets used and the labels contained within each dataset.

For the raw data, we removed duplicate entries and filtered out noisy samples such as meaningless word groups, and used regular expressions to eliminate special characters. Considering that large language models (LLMs) have limitations in processing information from the middle sections of long texts (Liu et al., 2024), we also removed text data that was either too long or too short. Finally,

LLM-Metaphor detection

1). few-shot:
Example 1:
Text: I have fallen into a deep well that I cannot come out of it, no matter how hard I try.
Answer:1(include metaphors)
Explain: Depressed patients believe that they are imprisoned in a prison of their depressive thoughts and conceptualize it as an impossible well to be released.
Example 2: ...
Example 3: ...
2). prompt:
Based on the example provided earlier, only determine whether the content after Post_text contains metaphorical parts, with 0 indicating not included and 1 indicating included. If there is a metaphor, describe the connection between metaphor and depression in one sentence. The answer format is as shown in the example.

Figure 2: Prompt for metaphorical annotation

LLM-Extract-source-domains

Prompt:
"Below are the source areas associated with depression, along with some reference definitions and examples from each area.
In context, identify and intercept depression-related metaphorical original texts from the following texts and select one or more suitable source domains from the list provided:
{domain_details}
Text:
{text}
The output format should be:
[{{"metaphor": "example metaphor", "source domain": ["domain1", "domain2"]}}]
or "non-metaphors" if no depression-related metaphors are found.

Figure 3: Prompt used for source domain annotation

the cleaned data was provided to the LLM for annotation.

3.2 Metaphorical Annotation

We used three large language models—Qwen-max, Llama-70B, and GPT-4—to perform binary annotation of metaphorical information on the cleaned dataset. The annotation was carried out using prompts as shown in the Fig.2, with the prompts consisting of two parts: [few-shot] and [prompt]. Since the goal was to analyze metaphors used by individuals with depression, the annotation process was conducted exclusively on the data labeled as depression-related within the dataset.

For the output results, we selected only the data where the annotation results were consistent across all three models. That is, we chose the data that all three models agreed contained metaphorical information, as well as the data that all three models agreed did not contain metaphorical information. For the data labeled as not containing metaphorical information, we conducted a manual review to ensure that these entries truly lacked metaphorical content. Ultimately, we obtained 6,119 entries containing metaphors, which will be used for further annotation, and 6,111 entries without metaphors.

3.3 Metaphor Source Domain Annotation

Based on the previous studies on depression metaphors and the proposed source domain classifications (Tonon, 2020;Coll-Florit et al., 2021;Barcelona Sánchez et al., 1986;Charteris-Black, 2012;McMullen and Conway, 2002), combined with an analysis of the obtained metaphor data, we have divided the metaphorical source domains related to depression into seven categories. For each metaphorical entry, we used GPT-4 to identify the corresponding source domain(s). GPT-4 was tasked with locating the metaphorical field within the data and selecting the appropriate source domain(s) from the seven predefined categories. A metaphorical field can be mapped to multiple source domains if applicable.

To guide GPT-4 in making the correct selections, we provided definitions and examples for each of the seven source domains, The final prompt is shown in Fig.3. These definitions and examples helped ensure that GPT-4 could accurately map the metaphors to their corresponding source domains, facilitating consistent and reliable annotation across the dataset. The seven source domain categories, their definitions, and examples are shown in Appendix A.1. In addition, we standardized the model's response format using prompts to facilitate subsequent manual review.

3.4 Manual evaluation

We used three groups of evaluators, each consisting of two people, to assess the model's selection of source domains. Initially, two groups performed the evaluation, and when their results disagreed, a third group decided which group's judgment was more appropriate. Before the evaluation, we introduced the evaluation criteria to the assessors and provided examples for demonstration. After training, the evaluators were randomly given 200 data points (Contains 384 metaphorical sentence-source domain pairs) to evaluate. The Kappa score (Fleiss, 1971) of the evaluation results $k \geq 0.8$, proving that our source domain classification is clear and reliable.

During the evaluation process, the evaluators primarily judged whether the metaphor fields identified by the model corresponded with the selected source domains. They retained the parts of the metaphor-source domain pair that matched in the sentences and removed the non-metaphor parts from the sentences. For parts where the metaphor field and the generated source domain did not match, the three groups discussed together to select the appropriate source domain for the sentence, aiming to improve the accuracy of the annotations.

"**text**": "Time to get biblical (rant : as a person who battles with demons everyday, I want to believe that in the end, our actions define us, not our urges or desires.)"

- "**label**": " depression "
- "**level**": "moderate"
- "**metaphor**": "contain-metaphor"
- "**source**": {
  "metaphor": "as a person who battles with demons everyday",
  "source domain": [" War ", "Living Organism"]
  }

Figure 4: An example of a metaphorical annotation

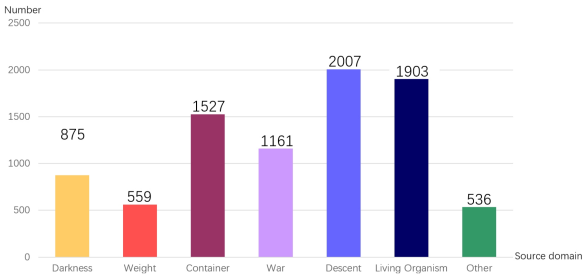

Figure 5: Number of each source domain

It is important to note that the evaluation only focused on the model's generated results; we did not assess other potential metaphor fields in the original text that the model failed to identify.

After evaluation, we finally obtained a dataset containing 4,426 data points related to depression source domains and 1,599 data points unrelated to depression metaphors. Since the number of non-depressed samples was noticeably smaller than the depressed samples, we also selected some tweets from non-depressed users in the MDDL dataset (Shen et al., 2017) to balance the ratio of depressed to non-depressed data. The final annotated dataset is illustrated in Fig.4.

4 Dataset Analysis

4.1 Depression Metaphor

The distribution of metaphorical information in the depressive samples within the dataset is presented in Tab.2. As can be observed, the proportion of users with depressive tendencies employing metaphors is approximately 50%. This figure is significantly higher than the proportion of metaphor usage in the daily language of the general public (around 32%) (Shutova and Teufel, 2010). This suggests that individuals with depressive tendencies may be more inclined to use metaphors when expressing emotions and experiences, possibly to convey complex inner feelings. Among the metaphors used by these individuals, those related to the seven concept source domains we defined account for the majority (73%), indicating that the classification of depressive source domains in this dataset is reasonable.

4.2 Source Domain Category

The distribution of samples across the seven source domains is shown in the Fig.5. The most common source domain is Descent, which indicates that a central experience of depression is the feeling of emotional decline and deterioration. The second most frequent source domain is Living Organism, where depression is often compared to an aggressive and destructive "living entity," such as a demon or a monster. The Container domain also appears more than 1,500 times, reflecting the sense of emotional confinement and suppression often experienced by individuals with depression. Apart from the Other category, Weight has the smallest proportion, suggesting that depression is more often perceived as a psychological experience rather than a physical one. Additionally, due to the individual differences among patients, the physical symptoms can vary. In the Other category, common metaphors include depicting depression as a journey or comparing it to a ticking bomb in one's mind. These distributions align in large part with previous studies on metaphors for depression.

Label	Number
depression-metaphors	6111
non-depression-metaphors	4426
depression-no-contain-metaphor	1599

Table 2: Number of metaphorical labels.

5 Experiment

We used the four types of labels included in the dataset—"Depression Status," "Depression Severity," "Presence of Metaphor," and "Metaphor Source Domain"—to construct a diverse set of instructions for fine-tuning the baseline model. This approach enhances the performance of the LLM in depression detection and enables the model to generate persuasive explanations for the detection results based on metaphorical information.

5.1 Experimental Setups

5.2 Instruction Construction

To enhance the model's generalization ability and practicality, ensuring it performs better in various contexts, we designed a hybrid selection instruc-

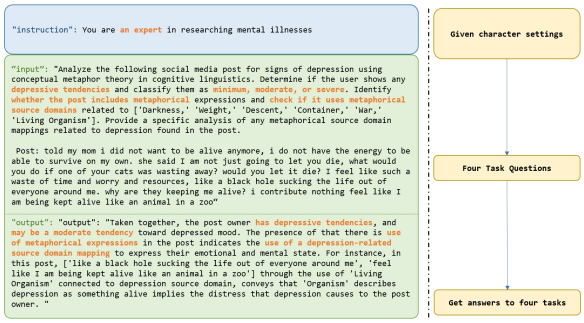

Figure 6: Examples of the constructed instruction set. The orange part in the Output will be replaced according to different labels.

tion scheme. We set up eight different instruction sets for eight sub-datasets, with each set containing five semantically similar but differently phrased instructions. This ensures that the labels of each sub-dataset are covered while randomly adjusting the positions of the labels. Each set of instructions was generated using templates and supplemented with five distinct expression variants generated by GPT, providing a diverse range of instruction formulations for each sub-dataset.

In terms of label processing, we designed five different expressions for the label of each task. For example, the label "depression" has five expressions: "has depressive tendencies", "has symptoms of depression", "has a certain degree of depressive trend", "is showing signs of depression", and "exhibits depressive tendencies". When constructing instructions, one of these expressions is randomly selected and added to the instructions. Finally, when generating instructions, we first randomly select different variants for all labels. Then, according to the task, we choose the corresponding template and insert the labels into the appropriate positions in the template. One example from the instruction set is shown in Fig.6.

By constructing a diversified instruction set, this design enables the model to more flexibly identify expression differences in depressive texts during mental health analysis. This enhances the model's ability to better support depression analysis and mental health intervention research. At the same time, it also helps to avoid the phenomenon of repetitive output that may occur after fine - tuning. Finally, we randomly mix the constructed instruction sets and divide them into a training set, a validation set, and a test set at a ratio of 8:1:1.

In this experiment, we selected Qwen2.5-7B (Yang et al., 2024a) and Llama 3.1-8B (Dubey et al., 2024) as the baseline models. After fine-tuning, both models were able to adapt well to domain-specific tasks and deliver high performance with relatively low resource requirements. We trained the models on two NVIDIA 4090 24GB GPUs. The learning rate was initialized to 1×10^{-4}, with 16-bit half-precision floating-point accuracy, and gradient accumulation steps were set to 8. Fine-tuning was conducted using the efficient model optimization tool LLaMA Factory (Zheng et al., 2024).

5.3 Results

5.3.1 Correctness

We evaluated the model's performance using F1 score, accuracy, recall, and comparisons with baseline models under zero-shot and few-shot conditions. The results, as shown in the Tab.3, demonstrate that the fine-tuned models achieved notable improvements across all three tasks. Specifically, the best-performing model, Llama-3.1-FT, showed approximately a 10% increase in F1 score for depression detection, a 27% improvement in depression severity detection, In the binary-classification metaphor detection task, as shown in Tab.4., the F1 scores of Llama3.1-FT and Qwen2.5-FT increased by approximately 32% and 14% respectively. This fully demonstrates the effectiveness of our dataset.

In depression detection task, compared to zero-shot, most models experienced a decrease in F1 scores under few-shot conditions, while their recall scores improved. This indicates that in few-shot settings, the models became better at accurately identifying positive examples but struggled more with distinguishing negative examples.

We also tested the model's ability in identifying depression-related source domains. Considering that the model's responses might contain descriptive statements rather than the specified source domain terms, we used semantic similarity metrics for evaluation, and the results are shown in the Tab.5. As can be seen, the fine-tuned model performs more accurately than the baseline model in identifying depression-related source domains, though further improvements are needed. We believe that future improvements could be made by refining the classification of depression-related source domains and by providing clearer definitions for these source domains.

5.3.2 Evaluation

We utilized diverse automated evaluation metrics to assess the model's performance in generating ex-

		Depression detection			Depression level recognition		
Model	**Parameter**	F1	Accuary	Recall	F1	Accuary	Recall
LLaMA3.1-ZS	8B	0.8793	0.8141	0.8559	0.3785	0.3932	0.3932
LLaMA3.1-FS		0.7578	0.6302	0.9151	0.3143	0.3778	0.3778
LLaMA3.1-FT		**0.9703**	**0.9704**	0.9655	**0.6486**	**0.6432**	**0.6432**
Qwen2.5-ZS	7B	0.7884	0.8174	0.6795	0.4623	0.5338	0.5338
Qwen2.5-FS		0.7282	0.6804	0.8554	0.4843	0.5000	0.5000
Qwen2.5-FT		0.8249	0.8499	0.7067	0.5193	0.6025	0.6025
Gemma-ZS		0.7480	0.7432	0.7717	0.4447	0.4478	0.4478
Gemma-FS		0.6641	0.4977	**0.9959**	0.3039	0.3247	0.3247
Mistral-ZS		0.7942	0.7998	0.7732	0.4347	0.4497	0.4497
Mistral-FS		0.6701	0.5245	0.9686	0.3217	0.3373	0.3373
LLAMA3.1-70B-ZS	70B	0.7920	0.8175	0.7035	0.5241	0.5200	0.5200
LLAMA3.1-70B-FS		0.7885	0.7980	0.7550	0.3402	0.3441	0.3441
GPT-4o-ZS	1800B	0.7773	0.812	0.6777	0.5689	0.5840	0.5840

Table 3: Performance of different models in depression detection and depression level recognition tasks."ZS" denotes zero-shot methods, and "FS" denotes few-shot methods."FT" denotes fine-tuning methods.

	Metaphor detection		
Model	F1	Accuracy	Recall
LLaMA3.1-ZS	0.6565	0.7025	0.5719
LLaMA3.1-FS	0.7292	0.7272	0.7388
LLaMA3.1-FT	**0.9757**	**0.9762**	0.9636
Qwen2.5-ZS	0.8313	0.8053	0.9653
Qwen2.5-FS	0.8537	0.8381	0.9504
Qwen2.5-FT	0.9715	0.9721	0.9587
Gemma-ZS	0.4146	0.4519	0.9936
Gemma-FS	0.4909	0.5827	**0.9959**
Mistral-ZS	0.5134	0.6398	0.9783
Mistral-FS	0.3879	0.3447	1.0000
LLAMA3.1-70B-ZS	0.7201	0.8536	0.9347
LLAMA3.1-70B-FS	0.7105	0.8387	0.9816
GPT-4o-ZS	0.6352	0.7680	0.9902

Table 4: Performance of different models in metaphor detection task.

planations. The results are shown in Tab.6. BLEU-4is a widely-adopted automated evaluation metric(Papineni et al., 2002), It calculates the precise match between the model's answers and reference answers based on the co-occurrence of 4-grams. ROUGE-L focuses on calculating the semantic consistency at the sentence or paragraph level(Lin and Hovy, 2003). Compared with traditional methods, BERTScore can capture more profound semantic information(Zhang et al., 2019), thus providing more accurate evaluation results in numerous cases. Inspired by MT-bench(Zheng et al., 2023), which uses LLM for generation quality evaluation, we also employed Llama-3.1-70b to score the model's generation quality. The scoring was carried out from three aspects: correctness, rationality, and conciseness, with scores ranging from 0 to 3. Correctness is to judge the degree of consistency between the model's answer and the reference answer. Rationality is to determine whether the model's answer is logical. Conciseness is to check whether the model's answer contains repetitive or redundant information. Finally, the average of the three scores was taken as the final result.

	Source Domain Identification		
Model	BLUE-4	ROUGE-1	BertScore
LLaMA3.1-ZS	19.75	22.43	81.74
LLaMA3.1-FS	19.27	28.11	81.66
LLaMA3.1-FT	**23.26**	**46.01**	**88.35**
Qwen2.5-ZS	13.28	23.71	81.11
Qwen2.5-FS	20.09	27.71	80.39
Qwen2.5-FT	31.5	43.67	88.05

Table 5: Performance of different models in a source domain identification task.

We calculated the average token length of the model's responses. As seen, compared to the baseline model, the fine-tuned Qwen2.5-FT is able to provide much shorter responses that are closer to the reference answers. In contrast, the baseline model's responses often contain a lot of irrelevant content or select incorrect conceptual source domains for metaphor fields, offering somewhat forced explanations. This results in lower evaluation scores for the answers from the baseline model. However, GPT-4o did not perform as expected. This is mainly because, in a zero-shot setting, GPT-4o tends to perform an extensive analysis of the post-content. Although the final conclusion is correct, the direction of its response differs significantly from the reference answer, leading to low BLUE-4 and ROUGE-L scores, but it performed

Model	BLUE-4	ROUGE-L	BertScore	MT-bench	Average token length
LLaMA3.1-ZS	0.1893	0.1078	0.8283	1.54	111.45
LLaMA3.1-FT	0.3844	0.3571	0.9052	2.1	**25.51**
Qwen2.5-ZS	0.1058	0.0781	0.8201	1.66	312
Qwen2.5-FT	**0.448**	**0.399**	**0.9105**	**2.33**	35.27
GPT-4o-ZS	0.0062	0.1361	0.847	1.94	205

Table 6: Evaluation of model generation results

	Depression detection			**Depression level recognition**		
Model	F1	Accuracy	Recall	F1	Accuracy	Recall
Qwen2.5	0.7281	0.7075	0.7823	36.36	29.65	29.65
Qwen2.5-FT						
+ (metaphor)	0.8187	0.8132	0.8422	**0.7496**	**73.02**	**73.02**
+ (metaphor + source)	0.9735	0.9741	0.9523	0.7425	72.45	72.45
+ (metaphor + source + source_location)	**0.9799**	**0.9803**	**0.9638**	0.7443	72.45	72.45

Table 7: Performance after constructing the instruction set fine-tuning model using different labels, with the labels used in constructing the instruction set in parentheses.

better on BERT-Score and MT-bench.

5.4 Supplementary experiment

To evaluate the role of the annotated metaphorical information in the dataset in improving the performance of LLM in depression detection, we designed a supplementary experiment. Although the fine-tuned Qwen2.5-7B performs worse than Llama-3.1-8B in depression detection, the performance of the Llama-3.1-8B baseline model is unstable. We found a large number of unavailable answers in the output of the Llama-3.1-8B baseline model. In such cases, the model indicates that it cannot provide analysis of depressive tendencies or mental health diagnosis information. This makes the performance of Llama-3.1-8B unstable, and similar situations may occur in other tasks. The proportions of similar situations in different tasks are shown in the Appendix A.2. We believe this is mainly due to the different training strategies adopted by different models.

Therefore, our supplementary experiment was conducted based on Qwen2.5-7B. We fine-tuned the model by constructing instruction sets using three different types of labels, and then tested the fine-tuned models' performance on two tasks: depression detection and depression severity detection. The experimental results are shown in Tab.7, indicate that, compared to the baseline model, using metaphor information effectively enhances the LLM's ability in depression detection. Furthermore, adding metaphor source domain information further improves the model's performance, while adding metaphor locating sentences provides limited improvement. In the depression severity detection task, the impact of metaphor information on the model is minimal, and the differences between the three fine-tuned models are not significant. We believe this is because the metaphors used by patients with varying degrees of depression are quite similar, and there is no evidence suggesting that the conceptual source domains they map to differ significantly.

6 Conclusion

In conclusion, this paper introduces a new public dataset for depression detection. This dataset includes the metaphors used by users with depressive tendencies and their corresponding source domain labels. Based on this dataset, we constructed an instruction set with diverse expressions to fine - tune two baseline models. The experimental results show that the fine - tuned models have significantly improved performance in multiple tasks. We also used automated evaluation metrics to assess the quality of the explanations generated by the models. The results indicate that the fine - tuned models can generate more concise, logical and higher - quality explanations. Our work demonstrates the potential of incorporating metaphorical information into automatic depression detection and explanation generation. In the future, we will further enhance the model's performance by refining the annotation of depression - related source domains and extend its capabilities to the detection of other mental illnesses.

Limitations

At present, one limitation of DRMD is that it only covers the English language. To promote comparative research on the commonly used metaphors of depressed patients across different language systems, it is highly necessary to develop DRMD datasets in other languages. Expanding with data from other languages will help identify the differences in metaphor usage among current different languages and deepen the understanding of the relationship between users' mental health status and the metaphors in their daily language. Moreover, datasets that include more diverse languages and cultures can further enhance the model's ability to understand metaphors. We encourage researchers to take on this challenging yet fascinating task by expanding DRMD through incorporating data from more languages in future work.

References

Antonio Barcelona Sánchez et al. 1986. On the concept of depression in american english: A cognitive approach.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023. Prompting for multimodal hateful meme classification. *arXiv preprint arXiv:2302.04156*.

Jonathan Charteris-Black. 2012. Shattering the bell jar: Metaphor, gender, and depression. *Metaphor and Symbol*, 27(3):199–216.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.

Marta Coll-Florit, Salvador Climent, Marco Sanfilippo, and Eulàlia Hernández-Encuentra. 2021. Metaphors of depression. studying first person accounts of life with depression published in blogs. *Metaphor and Symbol*, 36(1):1–19.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. 2019. Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 110–117.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

infamouscoder. 2022. Depression: Reddit Dataset (Cleaned). https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned/data.

Junhui Ji, Wei Ren, and Usman Naseem. 2023. Identifying creative harmful memes via prompt based approach. In *Proceedings of the ACM Web Conference 2023*, pages 3868–3872.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.

Richard R Kopp. 2013. *Metaphor therapy: Using client generated metaphors in psychotherapy*. Routledge.

Allison Lahnala, Yuntian Zhao, Charles Welch, Jonathan K Kummerfeld, Lawrence An, Kenneth Resnicow, Rada Mihalcea, and Verónica Pérez-Rosas. 2021. Exploring self-identified counseling expertise in online support forums. *arXiv preprint arXiv:2106.12976*.

George Lakoff and Mark Johnson. 1980. The metaphorical structure of the human conceptual system. *Cognitive science*, 4(2):195–208.

Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. 2020. Sensemood: depression detection on social media. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 407–411.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 150–157.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Dalia Magaa. 2019. Cultural competence and metaphor in mental healthcare interactions: A linguistic perspective. *Patient Education and Counseling*, 102(12):2192–2198.

Linda M McMullen and John B Conway. 2002. Conventional metaphors for depression. In *The verbal communication of emotions*, pages 167–181. Psychology Press.

Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. 2022. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022*, pages 2563–2572.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Anxo Pérez, Javier Parapar, Álvaro Barreiro, and Silvia Lopez-Larrosa. 2023. Bdi-sen: A sentence dataset for clinical symptoms of depression. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2996–3006.

Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task*, pages 9–12.

Kumarasan Roystonn, Wen Lin Teh, Ellaisha Samari, Laxman Cetty, Fiona Devi, Shazana Shahwan, Nisha Chandwani, and Mythily Subramaniam. 2021. Analysis and interpretation of metaphors: Exploring young adults' subjective experiences with depression. *Qualitative Health Research*, 31(8):1437–1447.

Kayalvizhi Sampath and Thenmozhi Durairaj. 2022. Data set creation and empirical analysis for detecting signs of depression from social media postings. In *International Conference on Computational Intelligence in Data Science*, pages 136–151. Springer.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, Wenwu Zhu, et al. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source-target domain mappings. In *LREC*, volume 2, pages 2–2. Citeseer.

Ellen Y Siegelman. 1993. *Metaphor and meaning in psychotherapy*. Guilford Press.

Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland. 2017. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web*, pages 715–724.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. *arXiv preprint arXiv:2106.01702*.

Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *Ieee Access*, 7:44883–44893.

Cristiana Tonon. 2020. Metaphors of psychological deterioration: the case of depression. *Le Simplegadi*, (20):122–135.

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Cheng Yang, Puli Chen, and Qingbao Huang. 2024b. Can chatgpt's performance be improved on verb metaphor detection tasks? bootstrapping and combining tacit knowledge. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1016–1027.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. 2022. D4: a chinese dialogue dataset for depression-diagnosis-oriented chat. *arXiv preprint arXiv:2205.11764*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2021. Depressionnet: A novel summarization boosted deep framework for depression detection on social media. *arXiv preprint arXiv:2105.10878*.

A Appendix

A.1 Classification of source domains

The definitions and examples of different source domains are shown in Tab.8

A.2 Unavailable responses

The proportions of the models generating unavailable responses in different tasks are shown in Tab.9

Source domain	Definition	Example
Darkness	This source domain portrays depression as a state conceptually dominated by dark colors.	1.A gray cloud is crying inside me. 2.It feels like a weight is on me at all times.
Weight	This type of metaphor described depression as whole body or spirit, or a particular part of the body, such as the head or heart feel heavy.	1.My feelings, far from improving, became very heavy. 2.I saw the whole world in black.
Container	This type of metaphor describes depression as a kind of conceptual bounded space, or the patient himself is some kind of conceptual container, the patient himself feels bound, oppressed, or abstractly unable to move	1.I locked myself inside myself. 2.We can't run away from ourselves.
War	This type of metaphor describes depression as an antagonistic force, similar to an opponent or enemy in battle.	1.I'm constantly fighting with my depression. 2.It s a war inside my head.
Descent	This type of metaphor describes depression as a process of conceptual downward progression, or the body, mind, or environment at some low point	1.I'm sinking deeper into despair. 2.I feel like I'm in hell.
Living Organism	This type of metaphor describes depression as something living, or something that has biological behavior.	1.It keeps growing inside me. 2.It' s like a parasite feeding on my joy.
Other	A sentence containing a metaphor related to depression, such as comparing depression to an abstract thing, a conceptual journey, or the patient likening themselves to something, and other sentences that do not belong to the other six metaphorical source domains.	1.My journey through depression. 2.I feel like little waves of panic wash through me all day.

Table 8: Source Domain Definitions and Examples.

	Depression detection	**Depression level recognition**	**Metaphor detection**
Model	False Ratio	False Ratio	False Ratio
LLaMA3.1-ZS	0.6968	0.6072	——
LLaMA3.1-FS	0.7931	0.9571	——
LLaMA3.1-FT-ZS	——	0.4442	——
Qwen2.5-ZS	——	——	——
Qwen2.5-FS	——	0.060	——
Qwen2.5-FT-ZS	——	0.0142	——
gemma-ZS	0.0781	0.3454	0.0127
gemma-FS	——	0.6323	——
mistral-ZS	——	——	0.1505
mistral-FS	——	0.0513	0.3652
llama3.1-70B-ZS	0.0160	0.0399	——
llama3.1-70B-FS	——	——	——
GPT-4o-ZS	——	——	——

Table 9: The proportion of unavailable responses of the models in different tasks,'——' indicates no unavailable responses