Application of Machine Learning Techniques for the Detection and Classification of Cognitive Distortions

Anonymous ACL submission

Abstract

Cognitive distortions are irrational or exaggerated thought patterns that negatively influence an individual's perception and emotional well-being. Accurate identification and classification of these distortions play a crucial role in supporting mental health diagnosis and therapy. In this study, we propose a machine learning-based approach to automatically detect and classify cognitive distortions from textual data. We construct a labeled dataset comprising various forms of distorted thinking, such as overgeneralization, catastrophization, and personalization, drawn from online forums and clinical sources. Using a combination of natural language processing techniques and supervised learning algorithms, we extract semantic features and train models capable of distinguishing between multiple types of distortions. Experimental results demonstrate the effectiveness of our approach, with notable improvements in precision and recall over traditional rule-based methods. This research lays the groundwork for the development of intelligent tools that can assist mental health professionals and enhance cognitive behavioral therapy through automated feedback and real-time distortion monitoring.

1 Introduction

Cognitive distortions refer to biased or irrational thought processes that magnify negative feelings and contribute to maladaptive behavior. These distorted patterns are frequently observed in individuals suffering from anxiety, depression, and other mental health disorders. In Cognitive Behavioral Therapy (CBT), one of the key therapeutic strategies is to help patients recognize and modify these thought patterns. While this approach is highly effective, identifying cognitive distortions in real-time typically requires skilled clinicians, creating a challenge when attempting to extend this practice to digital or automated systems. As the demand for scalable mental health support grows, there is a need for tools that can assist in the detection and intervention of cognitive distortions without relying solely on human therapists.

With the rise of online communication and the growing demand for mental health support, there is an increasing interest in developing automated tools to assist in early mental health interventions. Despite the emergence of several systems for emotion detection and sentiment analysis, relatively few studies have focused specifically on the automatic identification and classification of cognitive distortions in natural language, particularly from the speech or written reflections of patients.

This research aims to address this gap by leveraging modern Natural Language Processing (NLP) techniques, especially those based on Machine Learning (ML) and Deep Learning (DL), to detect and categorize different types of cognitive distortions in patient-generated language. Our main contributions are as follows:

- We explore the use of question-answering based models (e.g., Machine Reading Comprehension) to extract distortion spans from text.
- We implement a multi-class classification framework to categorize the type of cognitive distortion present in a given utterance.
- We analyze the effectiveness of various pretrained language models and fine-tuning strategies for both tasks.

By automating the detection and classification of cognitive distortions, our study aims to support the development of intelligent mental health assistants that can offer timely, low-cost, and scalable interventions.

2 Related Work

Cognitive distortions—irrational and negative thought patterns—are closely associated with mental health disorders such as depression and anxiety. The automatic detection and classification of such distortions have received growing attention in recent years, particularly with the rise of machine learning and natural language processing (NLP) techniques.

Shreevastava and Foltz (2021) were among the first to introduce a publicly available dataset focused on cognitive distortions in patient speech. They annotated 2,531 utterances with both binary (distorted or not) and multi-class (distortion type) labels. Their work proposed a semi-supervised learning framework that leveraged labeled and unlabeled examples to improve classification performance. Despite its contribution, the approach faced challenges with interpretability and generalizability due to the limited dataset size and lack of context.

Building on this foundation, Chen et al. (2023) explored the use of large language models (LLMs) such as ChatGPT and GPT-4 to detect and explain cognitive distortions in a more interpretable and human-aligned way. They introduced the Diagnosis of Thought (DoT) prompting framework, which guides the LLM through a three-step reasoning process: subjectivity assessment, contrastive reasoning, and schema analysis. Their results demonstrated that DoT prompting significantly outperformed zero-shot baselines and even matched or exceeded some supervised learning methods in both accuracy and interpretability. Furthermore, human evaluations by licensed psychotherapists confirmed the clinical usefulness of the generated rationales.

Together, these works highlight two major directions in cognitive distortion detection: supervised/semi-supervised learning based on annotated datasets, and prompt engineering leveraging the reasoning capabilities of large language models. Our work builds on these foundations by further exploring machine learning techniques for both detection and classification, while emphasizing performance, explainability, and clinical applicability.

3 Methodology

3.1 Problem Statement

Cognitive distortions are irrational and negatively biased thought patterns that can contribute to various mental health issues. Detecting and classifying these distortions in textual data is crucial for developing supportive tools in cognitive behavioral therapy (CBT). This study aims to leverage machine learning techniques to automatically identify and categorize cognitive distortions from patient-generated texts.

MRC-based Extraction of Cognitive Distortions Machine Reading Comprehension (MRC) models have shown significant promise in understanding and extracting relevant information from text. In the context of cognitive distortions, MRC can be utilized to pinpoint specific segments within a patient's narrative that indicate distorted thinking. By formulating the detection task as a question-answering problem, the model can be prompted with questions like, "What part of the text reflects a cognitive distortion?" This approach allows the model to focus on extracting evidence-based segments that signify distorted thoughts. Recent studies have demonstrated the efficacy of MRC frameworks in clinical concept extraction, highlighting their potential in identifying nuanced psychological patterns (Chen et al., 2023).

Cognitive Distortions Type Classification Once potential cognitive distortions are extracted, the next step involves classifying them into specific categories, such as overgeneralization, catastrophizing, or personalization. This classification task can be approached using supervised machine learning models trained on annotated datasets. Techniques like fine-tuned transformer-based models (e.g., BERT, RoBERTa) have been effective in capturing contextual nuances necessary for accurate classification. Incorporating structured prompting methods, such as the Diagnosis of Thought (DoT) framework, can further enhance the model's reasoning capabilities, leading to more interpretable and clinically relevant classifications (Chen et al., 2023). Additionally, leveraging datasets like the one introduced by Shreevastava and Foltz (2021), which contain expert-annotated examples of cognitive distortions, can significantly improve the model's performance and generalizability.

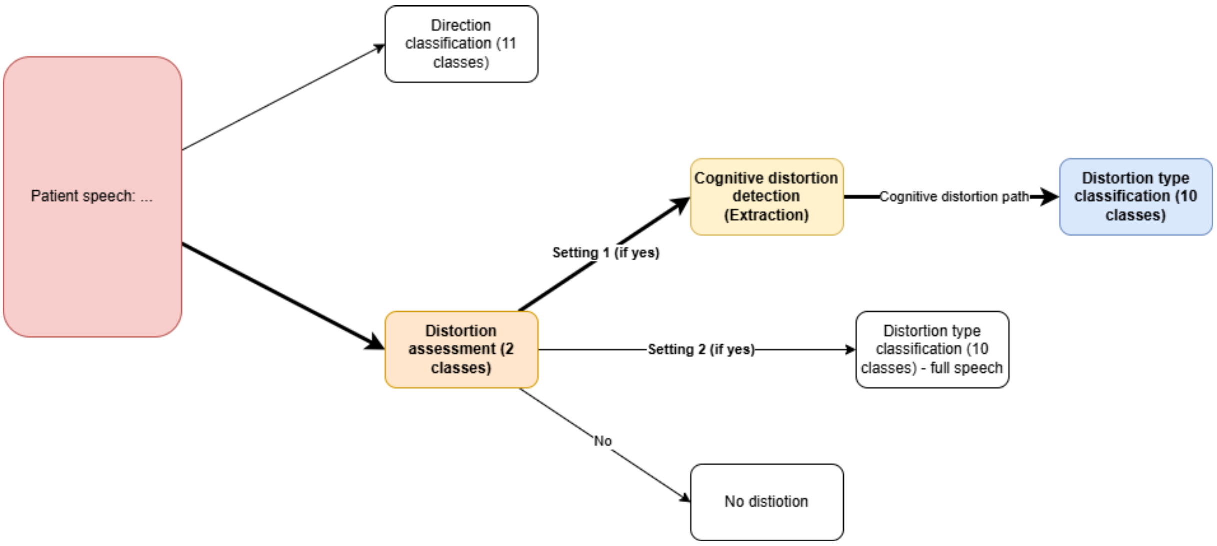
Direction classification (11 classes)
Patient speech: ...
Cognitive distortion detection (Extraction)
Cognitive distortion path
Distortion type classification (10 classes)
Setting 1 (if yes)
Distortion assessment (2 classes)
Setting 2 (if yes)
Distortion type classification (10 classes) - full speech
No
No distiotion

Figure 1: The proposed framework using cognitive distortion analysis for enhancing empathy in LLMs.

4 Experimental Settings

4.1 Datasets

4.2 Baselines

4.3 Settings

4.4 Evaluation metrics

5 Results and Discussion

5.1 Performance Comparison

MRC-based Extraction of Cognitive Distortions

Cognitive Distortions Type Classification

Limitations

Ethics Statement

Acknowledgements

References

Zhiyu Chen, Yujie Lu, and William Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304, Singapore. Association for Computational Linguistics.

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158, Online. Association for Computational Linguistics.

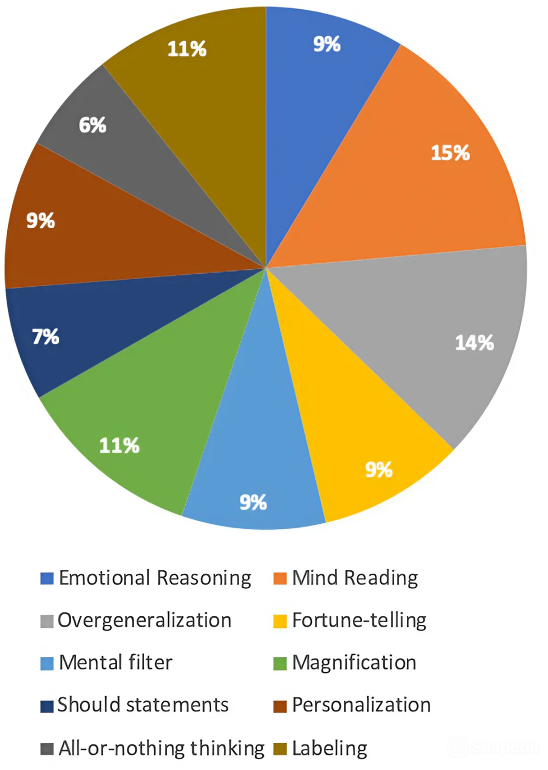

Figure 2: Distribution of the types of Cognitive Distortions(Shreevastava and Foltz, 2021).

A Example Appendix

This is a section in the appendix.