Abstract

The accelerating evolution of societal dynamics has brought forth an increasingly diverse array of information types. Of particular significance is the burgeoning interest in sentiment analysis, spurred by its versatile applications across various domains. The discernment of sentiments holds particular relevance, given its potential utility in numerous applications. Consequently, there has been a concerted effort to delve into the intricacies of multiple modalities to unearth latent information. This has given rise to a spectrum of methodologies aimed at effectively handling the complexities inherent in the amalgamation of diverse modalities. This has given rise to a spectrum of methodologies aimed at effectively handling the complexities inherent in the amalgamation of diverse modalities.

Concurrently, the societal discourse on mental health has manifested in an upsurge of applications pertaining to sentiment analysis and emotion detection. This evolving landscape has witnessed the trajectory of sentiment analysis tasks, progressing from unimodal and bimodal to the contemporary trimodal paradigm. The concomitant escalation in the demand for adeptly managing multiple modalities has been a discernible trend in recent years. Within this overarching milieu, this research introduces a text prompt-based fine-tuning method designed to address the challenges posed by distinct modalities within the framework of multimodal sentiment analysis.

The research objective is the pursuit of an interpretable and simplified approach for alleviate the gap between disparate modalities in a natural language manner. In this pursuit, an initial recourse is made to a prompt-based methodology during the fine-tuning phase. This methodological choice is grounded in its transformative capacity, recasting downstream tasks as cloze-filling exercises—a format inherently conducive to enhanced human comprehension. However, the matter lies in generating semantically rich representation from modalities beyond textual data.

To achieve this goal, a text prompt-based fine-tuning method is proposed in this research. This approach hinges on the meticulous application of manually crafted rules to generate textual descriptions from visual and auditory modalities. Consequently, the semantic descriptions is combined with textual information in a natural language fomrat with a fixed template. Due to its interpretability in natural language, this method is capable to understand by human beings. In other words, it also is able to make an adaption to different task. Subsequently, the process entails the formulation of a prompt function,

which is fed into a pre-trained language model and make the prediction. In the validation of this methodology, experiments are conducted leveraging the MELD dataset. Comparative analyses juxtaposing baseline results with an augmented baseline featuring attention mechanisms underscore the efficacy of the proposed method.

In conclusion, this research propose a method applying with the prompt-based fine-tuning method to navigate the intricate landscape of multimodal sentiment analysis. The fusion method between different modalities of interpretability and simplification is shown.