COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Minxue Niu

Modeling Affect in Speech and Language in the Presence of Natural Inconsistency

Tuesday, May 27, 2025
2:00pm – 4:00pm
3901Beyster / Hybrid – Zoom

ABSTRACT: Understanding human affect, including emotions and moods, is crucial for a wide range of applications, from improving mental health monitoring to enhancing user experiences in human-computer interaction. The ambiguous and subjective nature of human affect poses many challenges in developing affect models, such as disagreements in human labels and misaligned signals across modalities (e.g., text and voice). By addressing these challenges and effectively making use of the information embedded in these inconsistencies, we can work toward building more reliable, interpretable, and human-aligned affect recognition systems.

This dissertation investigates the causes and effects of inconsistencies in speech and language based affect models and explores strategies to mitigate their impact or to use them as informative signals. First, we examine **modality inconsistency**: emotions conveyed through text and vocal modalities do not always align. In the mental health domain, we demonstrate that their mismatch carries important information about people's mood and can serve as signals for mood disorder monitoring. More broadly, we show that modeling text and acoustic modalities separately in speech representation learning yields richer and more robust embeddings for various speech understanding tasks. Second, we examine **annotation inconsistency** observed in human labels, highlighting their sensitivity to annotation study designs. We compare human and Large Language Models (LLMs) generated annotations and find that LLMs achieve strong performance. Building on this insight, we propose integrating LLMs into the human annotation workflow, which improves both the annotator's experience and label quality. We then explore **interpersonal inconsistency**. We show that emotion labels significantly differ across demographic and personality groups, and incorporating annotator-level information can improve personalized speech emotion recognition models. Finally, we introduce a contrastive distillation framework that transfers LLMs' generalizable emotion knowledge into a lightweight model, which learns efficient, emotion-salient embeddings and can seamlessly handle unseen emotion label spaces without extra training.

Together, these findings offer a new perspective on the role of inconsistencies in affective modeling. It is important to understand their origins and implications, develop strategies to mitigate unintended ones, and leverage useful signals from them, to achieve a more comprehensive understanding and modeling of human affect.

CHAIR: Prof. Emily Mower Provost