Transformer-Based Detection of Elder Abuse and Neglect Risk in Reddit Dementia Caregiving Communities

Jessica A. Hernandez Chilatra, BSN[1] [2], Carolyn Pickering, PhD, RN[1], Xiaoqian Jiang, PhD[2]
[1]Cizik School of Nursing, University of Texas Health Science Center at Houston, TX, USA;
[2]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, TX, USA.

Introduction

Family caregiving involves high levels of stress that sometimes can escalate to abusive or neglectful instances[1]. The use of abusive and neglectful behaviors within this context is under-recognized and under-reported, and people living with dementia have a high incidence of abuse and neglect[2]. Online caregiver communities, such as Reddit forums, provide valuable insights into risks of elder abuse and neglect in dementia caregiving contexts. However, these narratives are unstructured, making it difficult to identify patterns of risk systematically.

Objective: To develop and evaluate an embedding-based NLP and machine learning pipeline that leverages transformer models to automatically detect signals of elder abuse and neglect in dementia-related caregiver posts from Reddit communities, with the long-term goal of supporting scalable early detection and intervention.

Data

Figure 1. Posts were collected from five dementia caregiver subreddits (n=3,629).

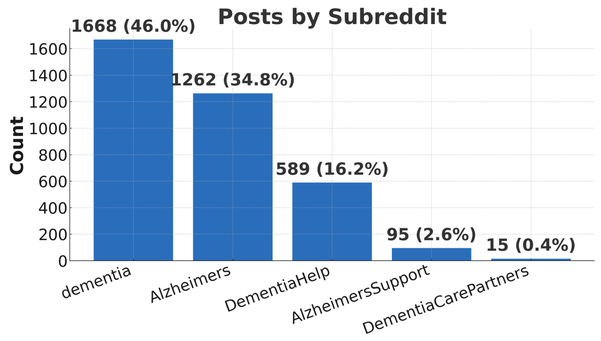
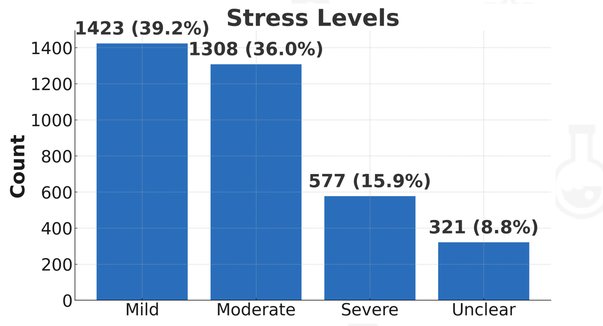
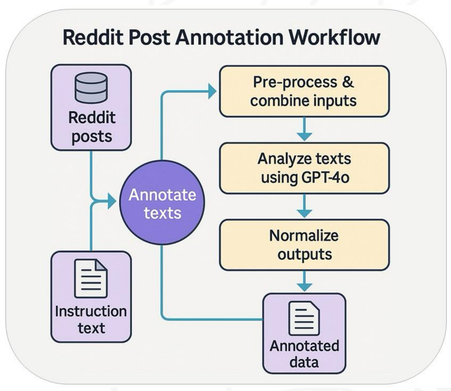


Figure 2. Most posts reflected mild to moderate stress; 16% showed severe indicators.



Annotation Pipeline & Modeling Approach

- **Annotation:** GPT-4o assisted + human review (~30%) using COPE Brief subscales, Zarit Burden Interview, and List of Stress Behaviors.
- **Model:** Fine-tuned MPNet to generate contextual embeddings → applied to supervised classifier (risk vs. no-risk of using abuse or neglect behaviors).



- **Split:** 70% train, 10% validation, 20% test.
- **Imbalance handling:** Weighted loss + threshold tuning (0.35–0.45 optimal).

Figure 3. Emotional and problem-focused coping were predominant, especially under severe stress.

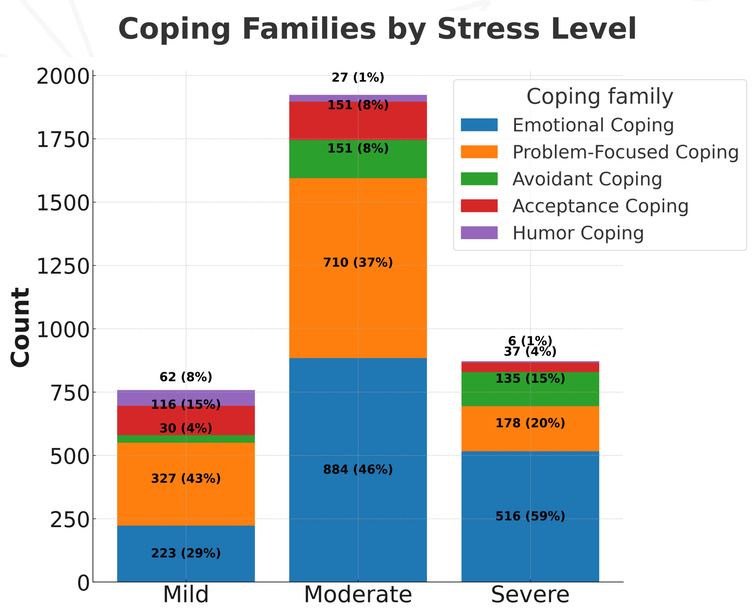
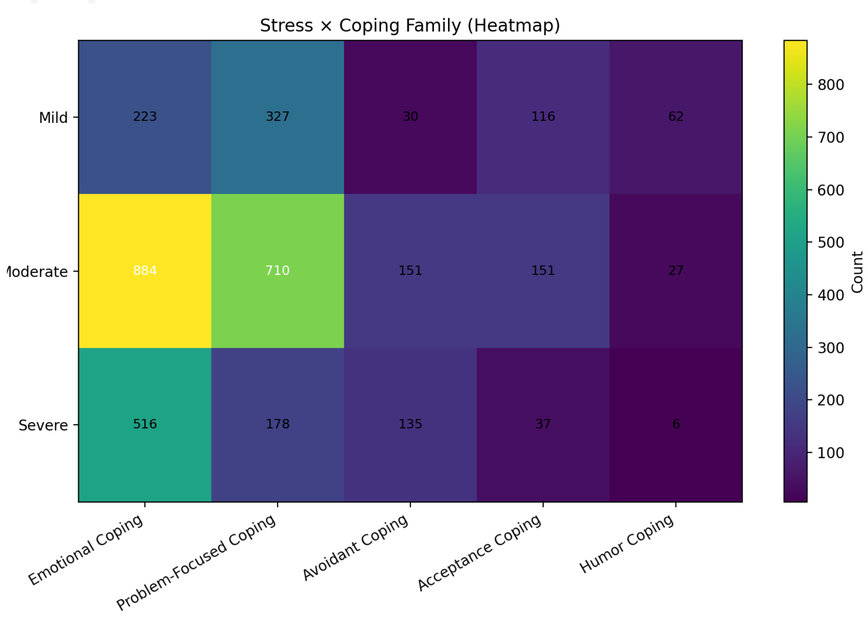


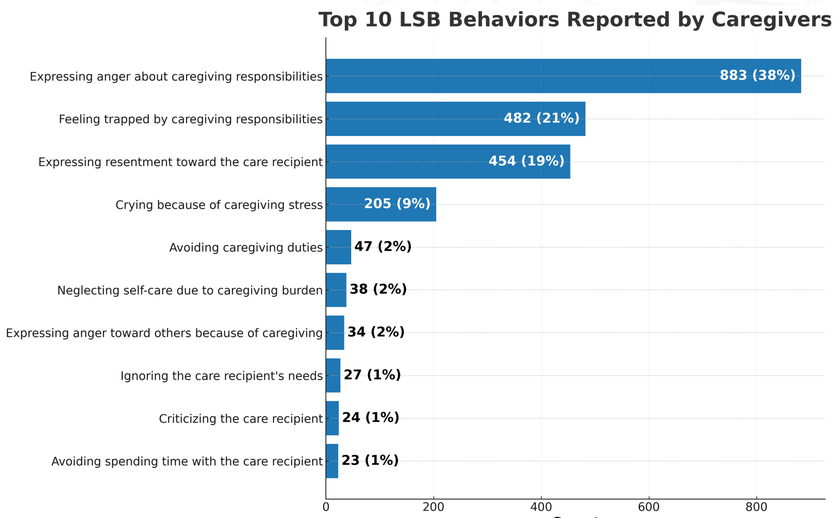
Figure 4. Visualizes proportional relationships between stress severity and coping type.



Results

Behavioral Indicators

Figure 5. Most Common Stress-Related Behaviors (n=4,764 total responses)

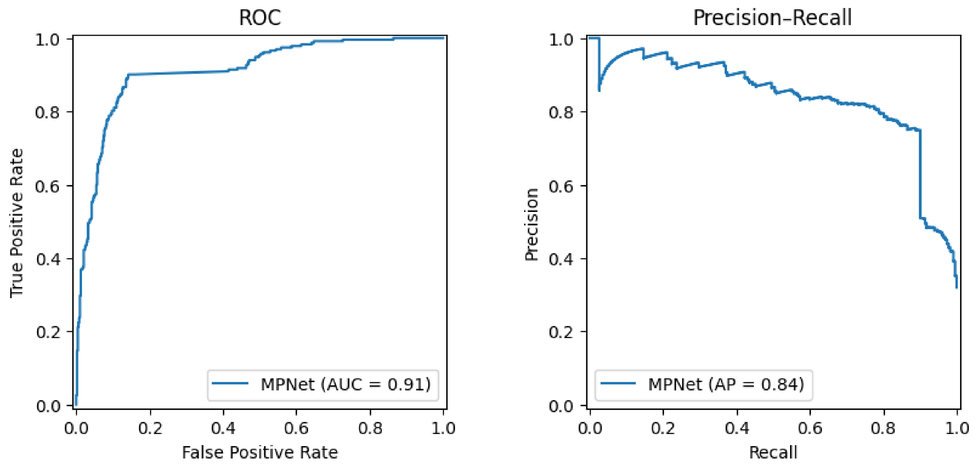


Model Results Summary

The fine-tuned MPNet classifier achieved:

- **Accuracy:** 86.6%
- **Precision**: 0.78
- **Recall**: 0.81
- **F1-score**: 0.79
- **ROC-AUC**: 0.91 | PR-AUC: 0.84
- **Brier Score:** 0.106 (after Platt scaling)

Figure 6. Model Performance Metrics.



Abuse/Neglect Prediction

Figure 7. Proportion of Posts Classified as At-Risk
The model identified 32% of posts as indicating risk of elder abuse or neglect.

Discussion & Implications

This work demonstrates that transformer-based NLP models can detect elder abuse and neglect risks from unstructured caregiver data. Results support the feasibility of scalable, AI-driven surveillance tools that complement existing screening methods and inform digital interventions.

Future directions:

- Integration with caregiver support platforms.
- Multilingual data expansion.
- Ethical safeguards and human-in-the-loop validation.

Acknowledgements

Supported by the University of Texas Health Science Center at Houston, McWilliams School of Biomedical Informatics, and Cizik School of Nursing.

Please contact the first author via email: Jessica.a.hernandezchilatra@uth.tmc.edu