Continually Adapting Language Models for Sailor Clinical Notes

Justin T. Stamets, MS[1,2]; Ryan W. Smith, PhD[1]; Andrew J. MacGregor, PhD[1]; Meg I. Robinson, MS[2 1]

[1]Naval Health Research Center, San Diego, CA; [2]Leidos, Inc., San Diego, CA

BACKGROUND

- Civilian and defense sectors have rapidly adopted Large Language Models (LLMs)
- Civilian sectors have seen a proliferation of domain-adapted models that are fine-tuned for specialized tasks
- While the United States Navy has pursued integrations with LLM providers, comparatively little research has been done to quantify the benefits of domain adaption specifically in the context of sailor health records

OBJECTIVES

1. Identify the limitations of relying on machine learning models that are unfamiliar with sailor health data
2. Leverage LLMs to model dynamic health trends in sailor datasets

METHODS

Study Population

- This was a retrospective analysis of provider notes from MHS Genesis, Theater Medical Data Store (TMDS), Periodic Health Assessments (PHA) between 2004–2025
- In total, a 46-million-word text corpus

Machine Learning Task and Evaluation Framework

- Topic modelling on held-out 2-million-word test set
- Success was quantified by topic coherence and topic diversity
- Scores are reported on evaluation of 25 runs with confidence intervals calculated from t-distribution margin of error

Machine Learning Models

- Latent Dirichlet Allocation (LDA) served as the statistical baseline
- Frozen ClinicalBERT within BERTopic framework served as the LLM approach
- Domain-adaptive pre-training (DAPT) of ClinicalBERT on 44-million-word training set served as the sailor model, which we call TotalSailorBERT

RESULTS: Model Performance

- LDA topic models were the weakest performer in both coherence (μ=0.530, 95% CI: 0.518-0.542) and diversity (μ=0.713, 95% CI: 0.686-0.740)
- Frozen ClinicalBERT within BERTopic significantly outperformed LDA in both topic coherence (μ=0.643, 95% CI: 0.626-0.660) and topic diversity (μ=0.875, 95% CI: 0.847-0.903)
- TotalSailorBERT achieved the similar topic coherence (μ=0.660, 95% CI: 0.646-0.674) to ClinicalBERT but was able to capture more nuance, evidenced by its topic diversity (μ=0.921, 95% CI: 0.907-0.935)

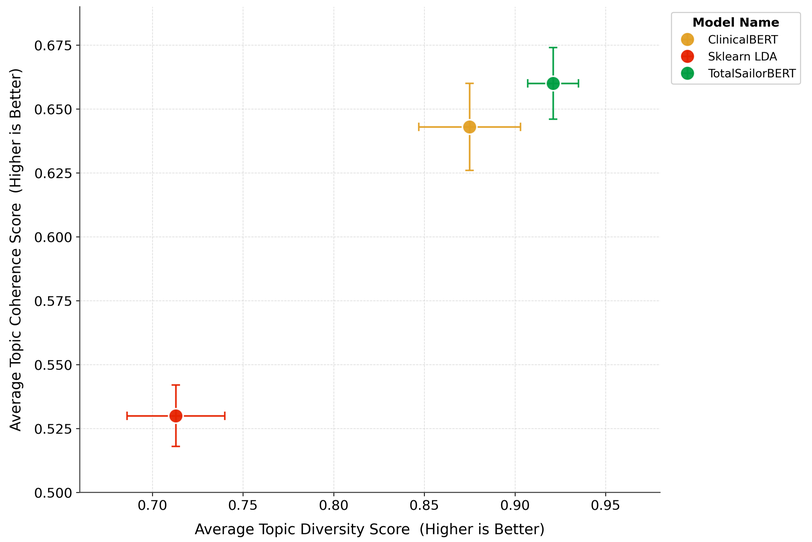

Figure 1. Model Performance: Topic coherence and diversity with confidence interval bands quantifying benefits of domain adaption

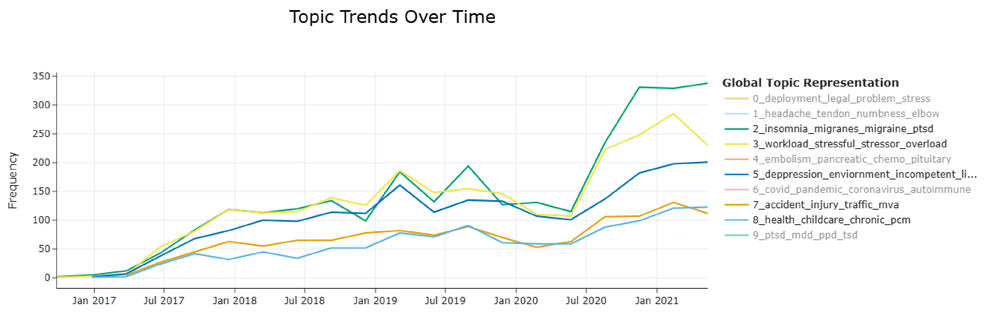

Figure 2. TotalSailorBERT Dynamic Topic Models. BERTopic models were used to analyze how self-reported stressors in the PHA evolved over time, revealing suggestive clinical patterns such as the intertwined dynamics of deployment stress and sleep issues.

RESULTS: Topical Content

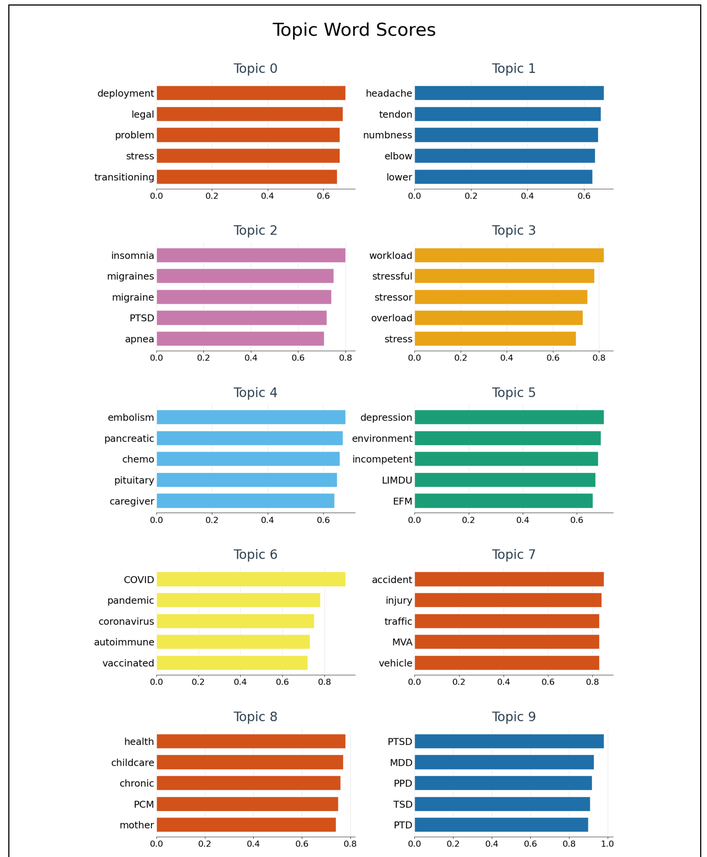

Figure 3. Topic Distributions: the topics learned by TotalSailorBERT are defined as a probability distribution over words. The mixture of this probability distribution defines general themes

CONCLUSIONS

- Pre-trained foundation models like ClinicalBERT provide strong general representations for clinical domains but can fail to capture dynamics that are more unique to sailors
- Techniques like DAPT enable LLMs to learn associations between general health topics and specific service member challenges, such as the relationship between sleep quality and deployment stress, which are critical for health surveillance and readiness modeling
- Future directions could incorporate these features within causal inference frameworks to enrich interventional capabilities with contextualized unstructured data points

 This work was supported by U.S. Navy Bureau of Medicine and Surgery under work unit no. 60808. The views expressed in this work are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government. The study protocol was approved by the Naval Health Research Center Institutional Review Board in compliance with all applicable federal regulations governing the protection of human subjects. Research data were derived from approved Naval Health Research Center Institutional Review Board protocol number NHRC.2003.0025.