

Assessing The Utility Of Unstructured Clinical Text For Dementia Risk Prediction Using Natural Language Processing



Anuththari Gamage and Roy Adams, PhD



Problem Statement

Pre-trained Language Models (e.g. BERT, Qwen, Llama) show success in risk prediction but interpreting black-box models is challenging.

Sparse Autoencoders (SAEs) combined with LLMs can be used to generate interpretations of transformer model internal mechanisms.

Generated interpretations are often highly predictive but lack grounding in domain knowledge – critical for clinical applications.

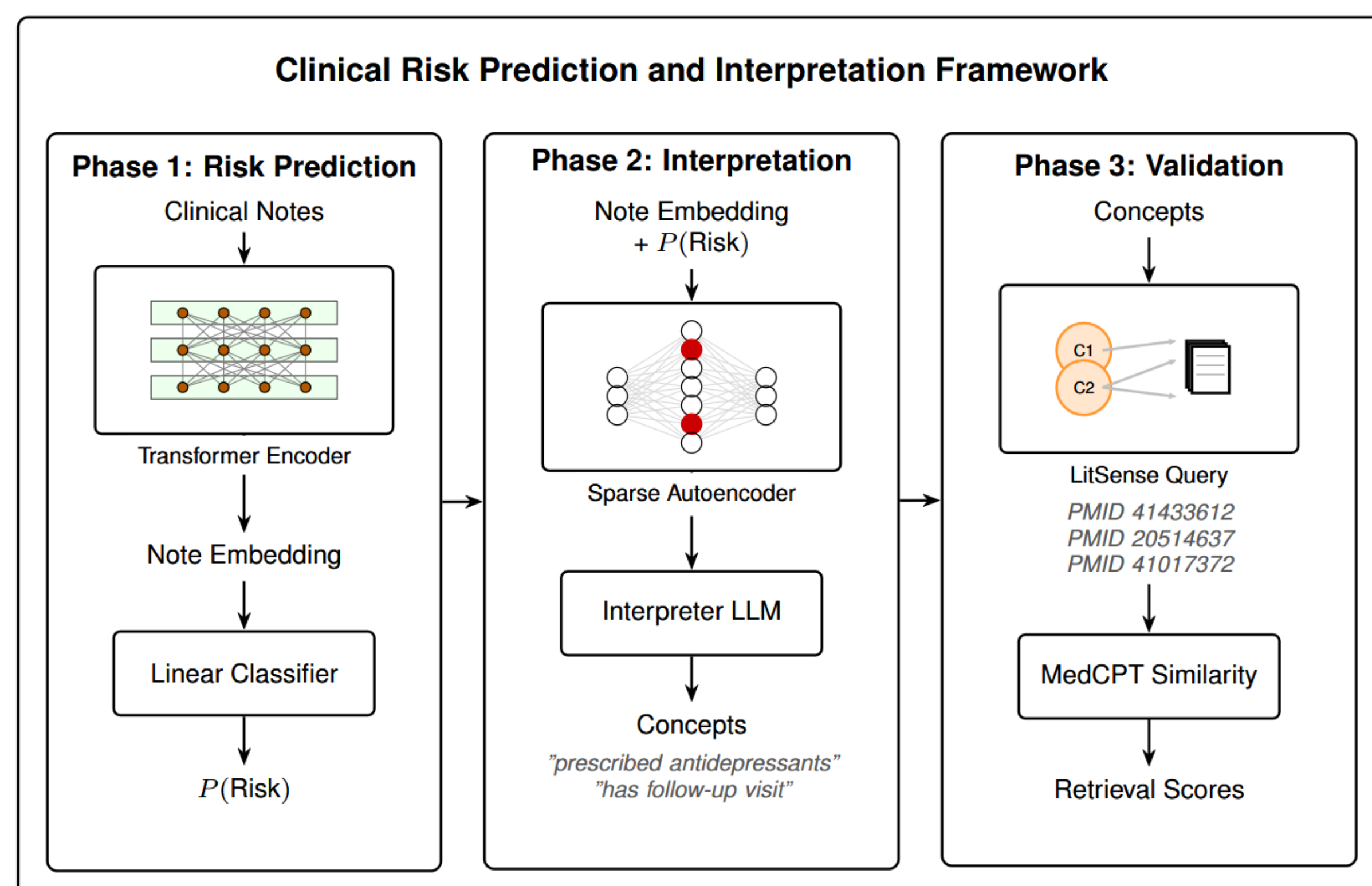
Methods

We adapted an SAE-based interpretation framework to risk prediction and extended it to check explanations against existing literature.

We evaluated model interpretations using three metrics: fidelity (alignment with model), predictivity (alignment with prediction target), and retrieval score (alignment with literature).

We applied this approach to dementia risk prediction using 10,046 notes from 4,300 patients at an urban academic memory clinic.

Framework



Predictive performance

Model	Context Window	Hidden Dimension	F1	AUC
Fine-tuned ModernBERT-large	8192	1024	0.576	0.755
Fine-tuned BioClinical ModernBERT-large	8192	1024	0.563	<u>0.745</u>
Fine-tuned BERT-large	512	1024	0.534	0.708
Fine-tuned AD-BERT	512	768	0.536	0.732
Fine-tuned BioClinical BERT	512	768	0.510	0.685
Fine-tuned BioBERT	512	768	0.530	0.703
Fine-tuned BERT	512	768	0.508	0.681
Bag-of-Words + Logistic Regression	-	-	<u>0.574</u>	0.730
Logistic Regression (structured data)	-	-	0.524	0.674

Model interpretations

#	Interpretations Green: + corr. Red: - corr.	Predictivity (Correlation)	Fidelity (F1)	Retrieval Score
15	patient has been previously diagnosed with MCI and is returning for a follow-up visit*	0.545	0.710	0.406
52	patient's spouse has taken on a caregiving or supervisory role due to patient's cognitive decline*	0.382	0.609	0.593
12	notes are follow-up evaluations with interval history sections authored by the same physician*	0.259	0.653	0
16	patient is prescribed or has tried donepezil (Aricept)	0.254	0.405	0.489
22	patient experienced post-operative delirium or surgery-related cognitive changes	0.210	0.471	1.000
60	patient experiences anxiety as a notable symptom	0.183	0.576	0.387
54	patient has been seen previously by the same clinician (follow-up visit with established patient)*	0.144	0.703	0.343
24	patient is currently prescribed an antidepressant medication (e.g., sertraline, venlafaxine, bupropion, or escitalopram) and their mood/depression status is explicitly discussed	-0.142	0.586	0.678
34	patients are middle-aged (roughly 50–58 years old) in positive samples vs. older patients (62–86 years old) in negative samples	-0.146	0.386	0.934
58	patient presents with cognitive complaints following a traumatic brain injury, cancer, or other non-neurodegenerative cause	-0.257	0.519	0.211
Overall predictivity (AUC) = 0.746				
*Significant hypotheses: 4/10 ($p < 0.01$)				