Improving the Generalizability of Natural Language Processing Algorithms in Medicine

Shreya Sripada[1], Kevin Xie[1], William K.S. Ojemann[1], Colin A. Ellis[1,2]

[1]Center for Neuroengineering and Therapeutics (CNT), [2]Department of Neurology, Perelman School of Medicine

ssripada@wharton.upenn.edu

- The Electronic Health Record (EHR) holds clinical information taken from the raw text of clinic notes written by healthcare providers
- Natural Language Processing (NLP) can be used to extract information out of this unstructured data
- However, these texts are vastly different: differing writing styles, medical jargon per specialty, and format
- We have previously explored Transformer-based models to extract outcomes from the clinic notes of patients with Epilepsy
- We explore similarity-based techniques taken prior literature [1] and how these generalize to other specialties

How do these techniques generalize to non-epileptologists notes?

How do similarity-based techniques compare to standard classification?

Methods

Gold Standard Annotations	Epileptologist (n=1,000, 700 training, 300 validation)*	Neurologist (n=100)	Non-neurologist (n=100)
Classification			
Seizure-Free	30%	33%	30%
Not seizure-free	62%	47%	35%
Unclassified	8%	20%	35%
Note contained seizure frequency	36%	14%	7%
Note contained date of most recent seizure	50%	48%	36%

The patients were classified as either Seizure Free (0), Has Seizures (1), or Could not classify (2).

Used Similarity-Based Techniques with different embeddings
- Lbl2Vec
- Lbl2TransformerVec (SimCSE)
- Lbl2TransformerVec (SBERT)
- Lbl2TransformerVec (SBERT 2)

Tested on three different testing sets
- Epileptologist notes (test)
- Neurologist notes
- Generalist notes

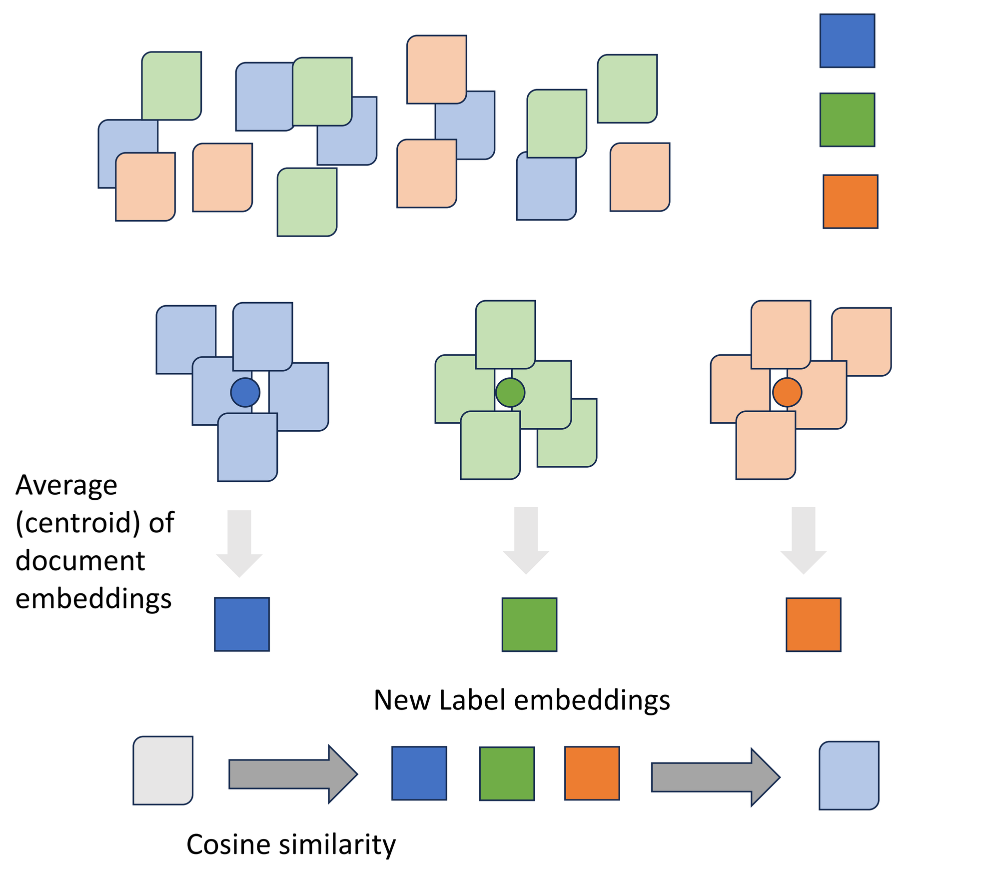
Training text + Label keywords used to train model

Model similarity of each test document to every label

Model uses similarity scores to classify each document to a label

Similarity Based Techniques

Embeddings: numerical representations of words, in the form of real-valued vectors



Label Keywords

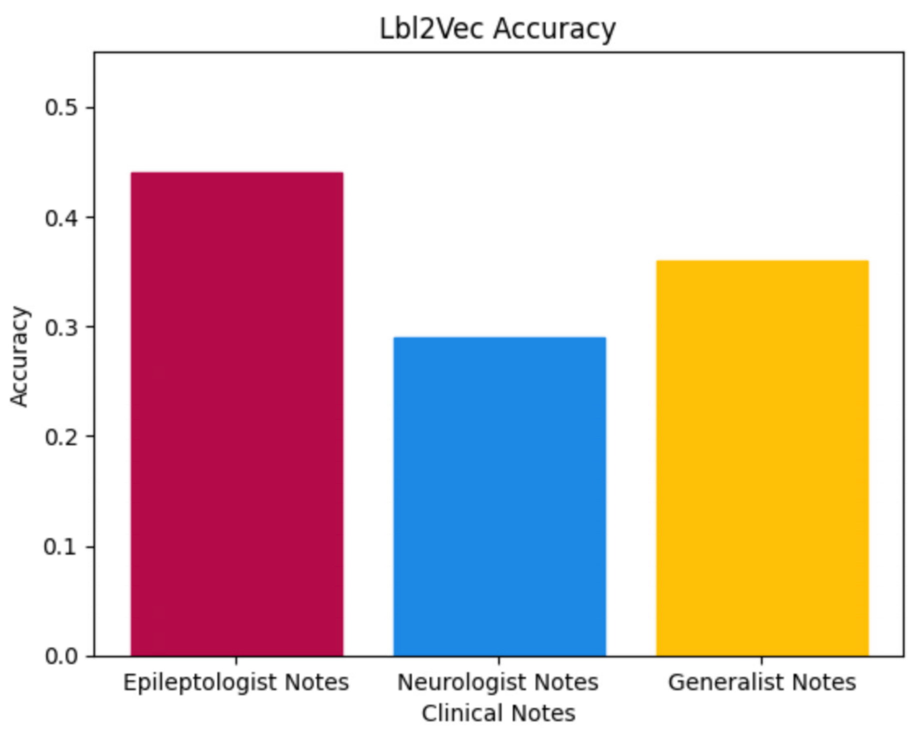
Seizure free keywords = ['seizure free', 'seizure stopped', 'denies seizures', 'no seizures', 'has not had seizures']
Has seizure keywords = ['had seizure', 'seizure relapse', 'seizure occurred', 'seizures recurring', 'remittent', 'abnormal movements', 'having convulsions', 'hands shaking', 'confused']
Unknown keywords = ['unknown', 'not', 'classify', 'unclear', 'last seen', 'stable']

Seizure free keywords = ['seizure', 'free', 'none', 'stopped']
Has seizure keywords = ['seizure', 'relapse', 'occur', 'recurring', 'having', remittent']
Unknown keywords = ['unknown', 'not', 'classify']

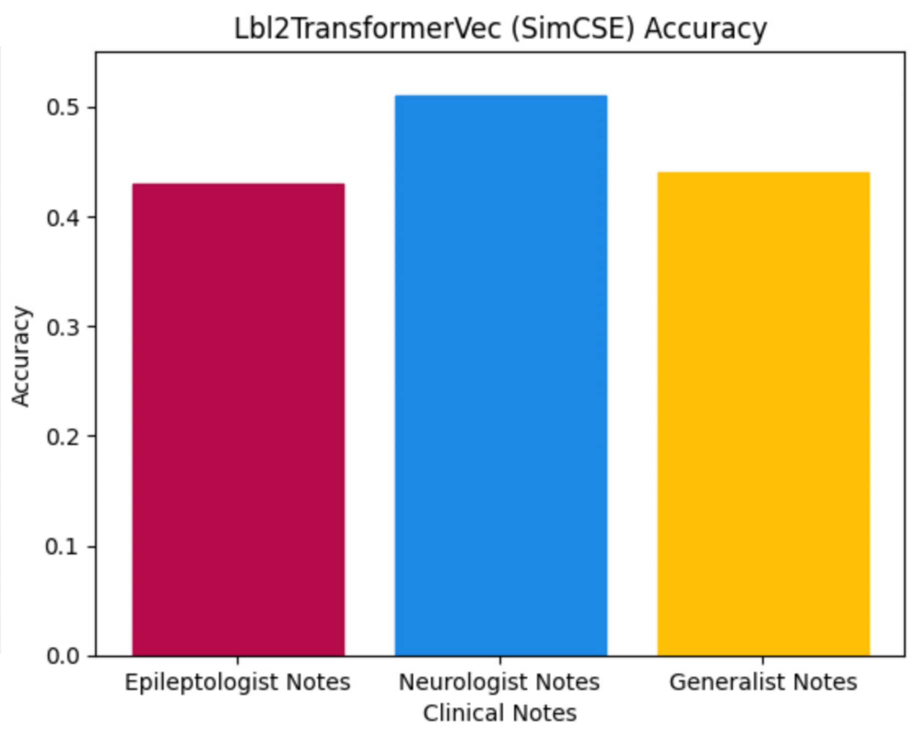
Results

Similarity Based Techniques

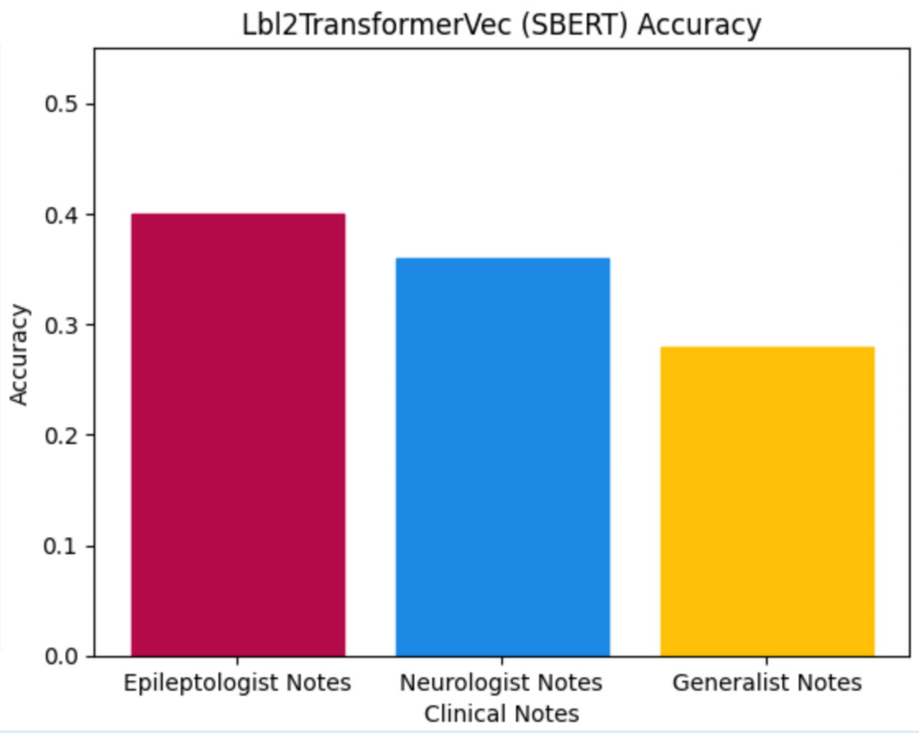
Lbl2Vec



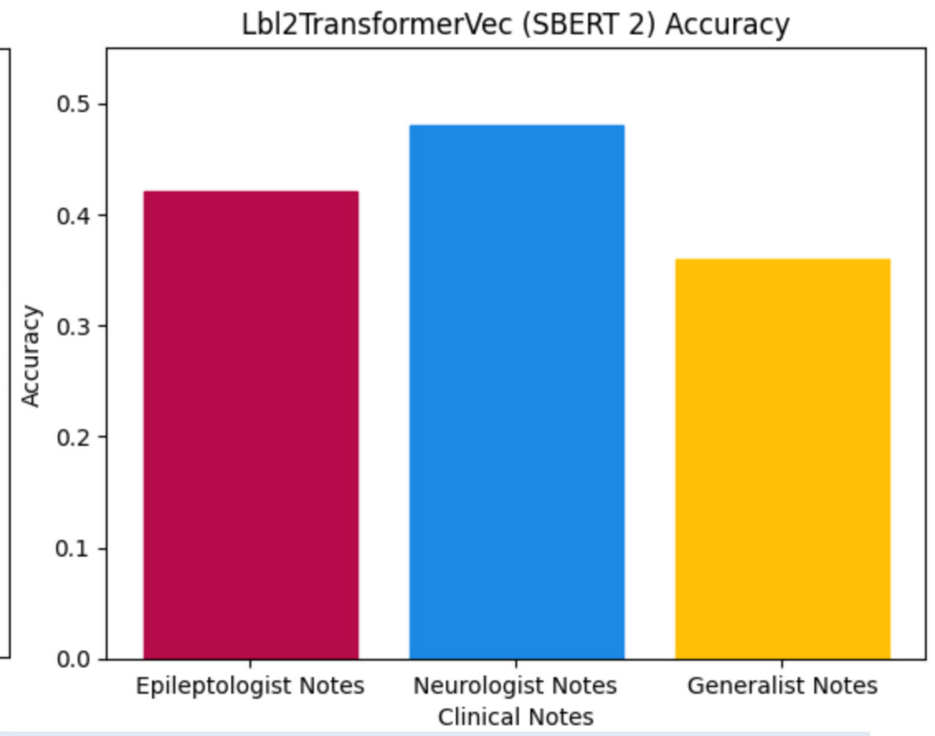
Lbl2TransformerVec (SimCSE)



Lbl2TransformerVec (SBERT)



Lbl2TransformerVec (SBERT 2)

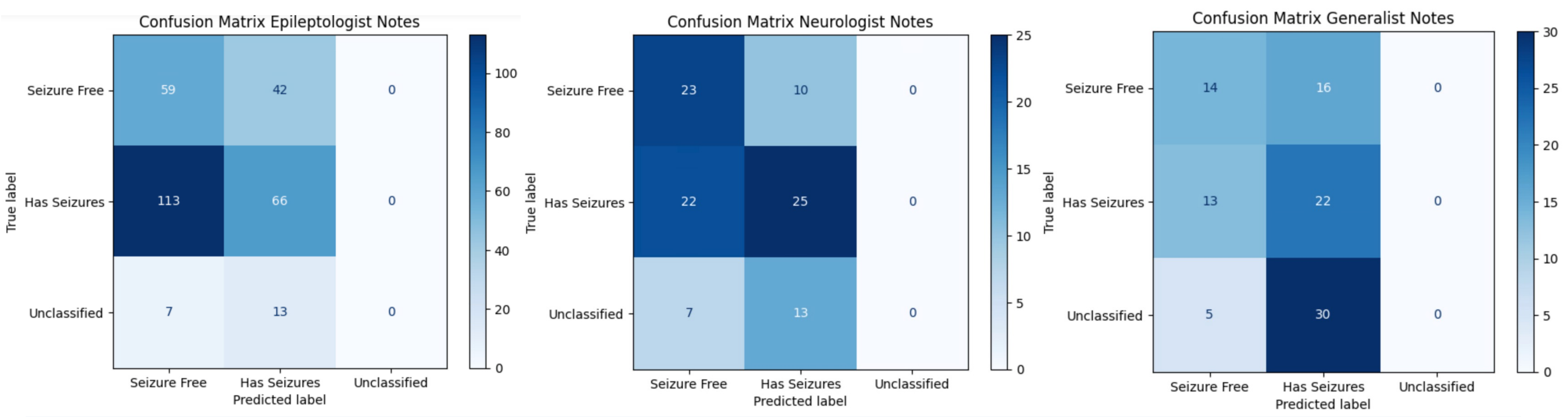


Used words as label keywords for the Lbl2Vec as the embedding baseline is Word2Vec (word-based)

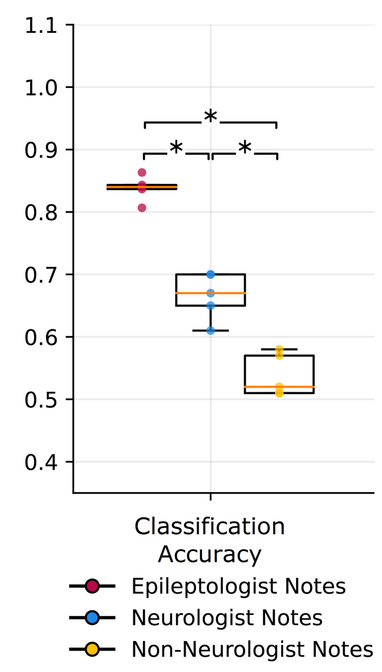
Used phrases as label-keywords for the transformer similarity-based techniques. The embedding baseline for these techniques are sentence-based techniques.

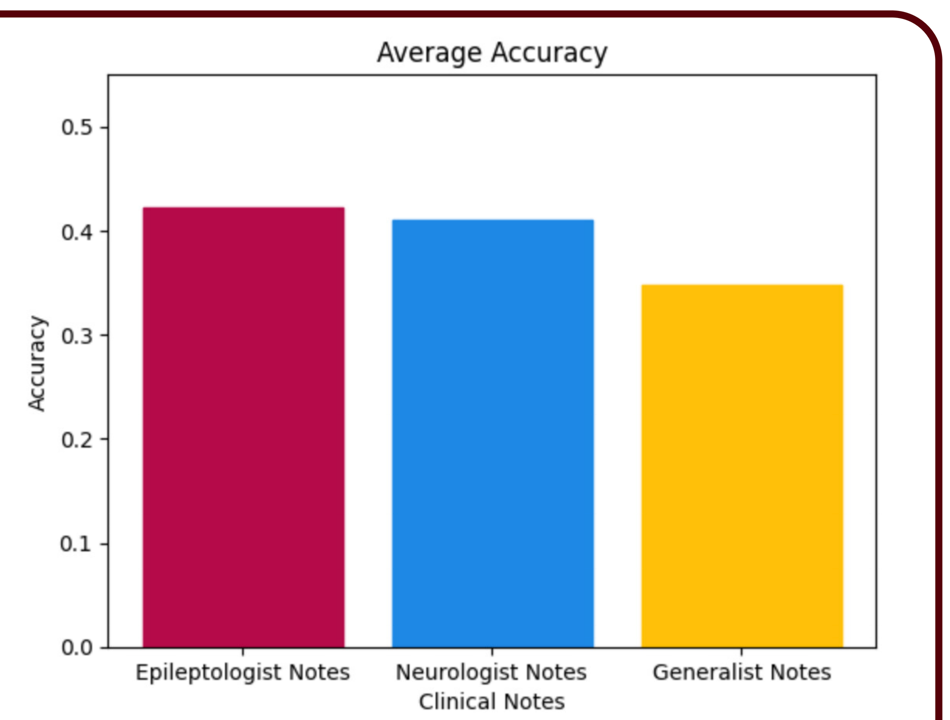
Not much consistency between techniques. Generalist notes seems to perform the worst. Neurologist more generalizable. Lbl2TransformerVec (SimCSE) performed the best.

Confusion Matrices



Model did not predict any as "Unclassified" (2). The label keywords are not specific and therefore the model does not perform well.





Epileptologist notes performed the best as expected. Similarity-based techniques did not perform as well as the standard classification [2].

Conclusions

Findings
- Similarity-based techniques perform better when it is the binary classification of "seizure free" or "has seizures" because the keywords are specific.
- The models are better at classifying epileptologists notes when trained on epileptologist notes, but the worst at generalist notes
- There is not much consistency between varying Lbl2Vec techniques
- There is room for improvement in identifying keywords

Limitations
- Our approach is agnostic to the type of seizure and provides only one of each outcome measure per note, potentially missing information
  - Seizures have varying severities and our NLP algorithms cannot account for or classifying that at this point
- Our notes and models were affected by copy-forwarded information, where a note author will copy previous notes into the current note, potentially introducing outdated/contradictory information.

Acknowledgements
- CNT Staff and Data Managers

References: [1] T. Schopf *et al.*, *NLPIR*, Dec. 2022; [2] K. Xie et al., *Journal of the American Medical Informatics Association.*, May. 2022