Improving the Identification of Substance Use from Clinical Notes with BERT

Stuart J. Waller BS, Cosmin A. Bejan PhD
Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

Introduction

Social determinants of health (SDoH) – the social and economic conditions that influence health risks and outcomes – have a significant effect on overall well-being across a lifespan. Addressing the underlying factors associated with SDoH is essential towards improving health and reducing health disparities. Despite recent efforts to better integrate SDoH into electronic health records (EHRs), there's no standardized framework to automatically capture this information. Motivated by the recent success of the Bidirectional Encoder Representations from Transformers (BERT) language model in various natural language processing (NLP) applications, we evaluated 5 BERT models on extracting the status of 3 categories of substance use (tobacco, alcohol, and drug) from clinical text.

Methods

Our dataset contained 2,220 de-identified clinical notes from the Vanderbilt University Medical Center's EHR data. It was manually annotated with drug use information, extending our previous annotations on tobacco and alcohol. The annotations were performed at mention-level and consisted of 6 categories for tobacco use (*Current Smoker*, *Past Smoker*, *Never Smoker*, *Unknown Smoker*, *Smoker*, and *Secondary Smoker*), 5 similar categories for alcohol use, and 5 similar categories for drug use. We associated each note with an *Ever*/*Never* binary category: *Never* and *Unknown* were mapped to *Never User* (NU) while the rest of the categories constituted *Ever User* (EU). Tobacco had 420 EU and 1,800 NU, alcohol had 663 EU and 1,557 NU, and drug had 49 EU and 2,171 NU.

Leveraging the *Ever*/*Never* categories, we trained a binary classification model for each substance – as our goal was to identify patients with substance use phenotypes across the entire EHR. Each note was preprocessed using an NLP pipeline: text was set to lowercase, and punctuation and symbols were removed. Because BERT models allow input sequences of at most 512 tokens, we also employed a keyword-based truncation approach (our average untruncated note was 1,022 tokens). For each substance, we identified a list of relevant keywords and truncated each note around the first detected keyword.

In addition to vanilla BERT-Large, we selected 4 state-of-the-art BERT models with the largest available sizes. We used PubMedBERT and BioELECTRA, both pretrained on biomedical domain text. ELECTRA employs a different pretraining technique and ConvBERT utilizes a different model architecture. We also tried ClinicalBERT, – a model pretrained on EHR data, – but the performance was significantly worse. Each of our 5 models was fine-tuned on the training set of each substance's truncated dataset and evaluated on the corresponding test set. We reported the binary classification performances on the test set of each substance's truncated dataset using precision (P), recall (R), and F1 score (F1).

Results

ELECTRA-Large achieved the best performance on tobacco across the board (95.6% P, 93.0% R, 94.1% F1) as well as the highest recall (98.1%) and F1 (94.3%) for alcohol (Table 1). ConvBERT-Base achieved a comparable F1 (94.1%) and the highest precision value (92.7%) for alcohol. All 5 models yielded similar drug scores with the exception of BERT-Large's distinguishably high F1 (94.1%). PubMedBERT was the only model that didn't produce a higher F1 for alcohol than tobacco.

Table 1. Evaluation of 5 BERT models on substance use extraction.

Model	**Tobacco Use**			**Alcohol Use**			**Drug Use**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BERT-L	94.0	91.4	92.7	89.3	97.2	92.8	88.9	**100**	**94.1**
ELECTRA-L	**95.6**	**93.0**	**94.1**	90.6	**98.1**	**94.3**	88.9	93.8	93.3
ConvBERT-B	95.3	91.4	92.9	**92.7**	96.3	94.1	83.8	93.8	88.9
PubMedBERT	95.0	92.0	93.2	90.9	97.2	93.1	88.9	**100**	93.3
BioELECTRA	94.8	90.3	92.4	90.6	94.9	92.6	**100**	93.8	93.3

Conclusions

Our results suggest that state-of-the-art language models are effective for extracting SDoH information from EHR data. Future work includes identifying more specific substance use phenotypes (e.g., *Type*, *Amount*, and *Frequency*) and constructing a longitudinal profile of these phenotypes at patient-level.