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 health and 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 is no standardized framework to automatically capture this information. Nevertheless, health care providers commonly describe such determinants in clinical notes. 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[1].

Methods

Our dataset includes 2,220 de-identified clinical notes from the Vanderbilt University Medical Center's EHR data. This dataset was manually annotated with drug use information, extending our previous annotation efforts on tobacco and alcohol use[2]. The initial 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 categories for alcohol use (*Current Drinker*, *Past Drinker*, *Never Drinker*, *Unknown Drinker*, and *Drinker*), and 5 status categories for drug use (*Current Drug User*, *Past Drug User*, *Never Drug User*, *Unknown Drug User*, and *Drug User*). The mention-level annotations were then converted to note-level into *Ever*/*Never* binary categories as follows: *Never* and *Unknown* were mapped to *Never User* while the rest of the categories constituted *Ever User*. The number of positive and negative samples for each substance in presented in Table 1.

The task we proposed to solve was a binary classification problem with the goal of automatically assigning an *Ever*/*Never* category for each note and substance use determinant. For this, all 2,220 notes were preprocessed via an NLP pipeline: text was set to lowercase, punctuation and symbols were removed, and sequences of 3 or more repeating characters were also discarded. To address the primary limitation of BERT models, which allow input sequences of at most 512 tokens, we employed a keyword, rule-based truncation approach (the average note length of our dataset was 1,022 tokens). For each determinant, we identified a list of relevant keywords (e.g., the keyword set for drug use included *marijuana*, *cocaine*, *crystal meth*, *xanax*, *weed*, and *drug*). Each note was truncated to 35 characters before the presence of the first detected keyword in a note and 165 characters after. If a note didn't contain any substance-specific keywords, the same truncation parameters would be applied to the designated fallback *social history* keyword as the social history portion of each note was explicitly marked.

In addition to vanilla BERT, we selected 4 state-of-the-art BERT models, each possessing a unique modification to give us a broad scope of capabilities. For our binary classification task, we added a single classification layer to each BERT model. PubMedBERT and BioELECTRA are pretrained from scratch on PubMed abstracts[3, 4]. ELECTRA employs a different pretraining technique (*replaced token detection* instead of *masked language modeling*) and ConvBERT utilizes a different model architecture (*span-based dynamic convolution* instead of *self-attention heads*) [5, 6]. For each BERT, we used the largest model size available as BERT models with more parameters tend to outperform their less expensive counterparts. BERT-Large, for example, significantly outperforms BERT-Base across all tasks, especially those with very little training data[1]. The full breadth of model sizes and pretraining details are presented in Table 2.

Table 1. Distribution of *Ever*/*Never* categories for each substance use phenotype.

SDoH	Ever User	Never User
Tobacco	420	1800
Alcohol	663	1557
Drug	49	2171

Table 2. Pretraining details of 5 BERT models. L, Large; B, Base

Models	Parameters	Corpus	Text Size
BERT-L	~340M	Wiki + Books	3.3B words
ELECTRA-L	~340M	Wiki + Books	3.3B words
ConvBERT-B	~110M	Wiki + Books	3.3B words
PubMedBERT	~110M	PubMed	3.1B words
BioELECTRA	~110M	PubMed	4.2B words

We split each of the substance-specific truncated datasets into a train, validation, and test set using a 70/10/20 ratio. The stochastic nature of fine-tuned machine learning models can yield significant variation in performance on small datasets depending on the random seed. Therefore, we report the median of 10 fine-tuning runs, each with a different random seed, to combat this variability. For all 5 BERT models, we experiment with the following hyperparameters: learning rate [2e-5, 5e-5], epochs [12, 6, 5], and batch size [32]. A grid search was performed on the validation set to find an optimal set of hyperparameter values. We record the binary classification performances on the test sets using precision, recall, and F1-score.

Results

As listed in Table 3, ELECTRA-Large achieved the best performance on tobacco across the board (95.6% P, 93.0% R, 94.1% F1) in addition to recall (98.1%) and F1-score (94.3%) for alcohol. ConvBERT-Base saw a comparable F1-score (94.1%) and the highest precision value (92.7%) for alcohol. Due to the small amount of positive drug samples, all 5 models yielded similar scores with the exception of BERT-Large's distinguishably high F1-score (94.1%). Furthermore, PubMedBERT was the only model that didn't produce a higher F1-score for alcohol than tobacco.

Table 3. Evaluation of 5 BERT models on substance use extraction from clinical notes.

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

P, Precision; R, Recall; F1, F1-score

Conclusions

This study depicts the evaluation of 5 BERT models on extracting 3 substance use phenotypes from clinical notes. Our results suggest that state-of-the-art language models are effective for extracting SDoH information from EHR data. Future work includes extracting more specific substance use phenotypes (such as *Type*, *Amount*, and *Frequency*) as well as constructing a longitudinal profile of these phenotypes at patient-level.

References

1. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171-4186, 2019.
2. Chen J, Ermias A, Bejan C. Using clinical notes to assess the exposure of tobacco and alcohol use in the electronic health record. 2019.
3. Yu G, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-specific language model pretraining for biomedical natural language processing. 2020.
4. Kanakarajan K, Kundumani B, Sankarasubbu M. BioELECTRA: Pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143-154, 2021.
5. Clark K, Luong M-T, Le QV, Manning CD. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2020.
6. Jiang Z, Yu W, Zhou D, Chen Y, Feng J, Yan S. ConvBERT: Improving BERT with span-based dynamic convolution. In *Advances in Neural Information Processing Systems*, 33, 2020.