CASE: Efficient Curricular Data Pre-training for Building Assistive Psychology Expert Models

Sarthak Harne*, Monjoy Narayan Choudhury*, Madhav Rao, TK Srikanth, Seema Mehrotra, Apoorva Vashisht, Aarushi Basu, Manjit Sodhi

*equal contribution

Data Scarcity for Training Models

1. BERT is trained on the entirety of English Wikipedia (~2.5B words) and Google's BooksCorpus (~800M words) and has good syntactic knowledge of Language, but not domain specific semantic knowledge.

2. Further pretraining and fine tuning is required for this. However, for pretraining we current state-of-the-art methods use similar scale of data.

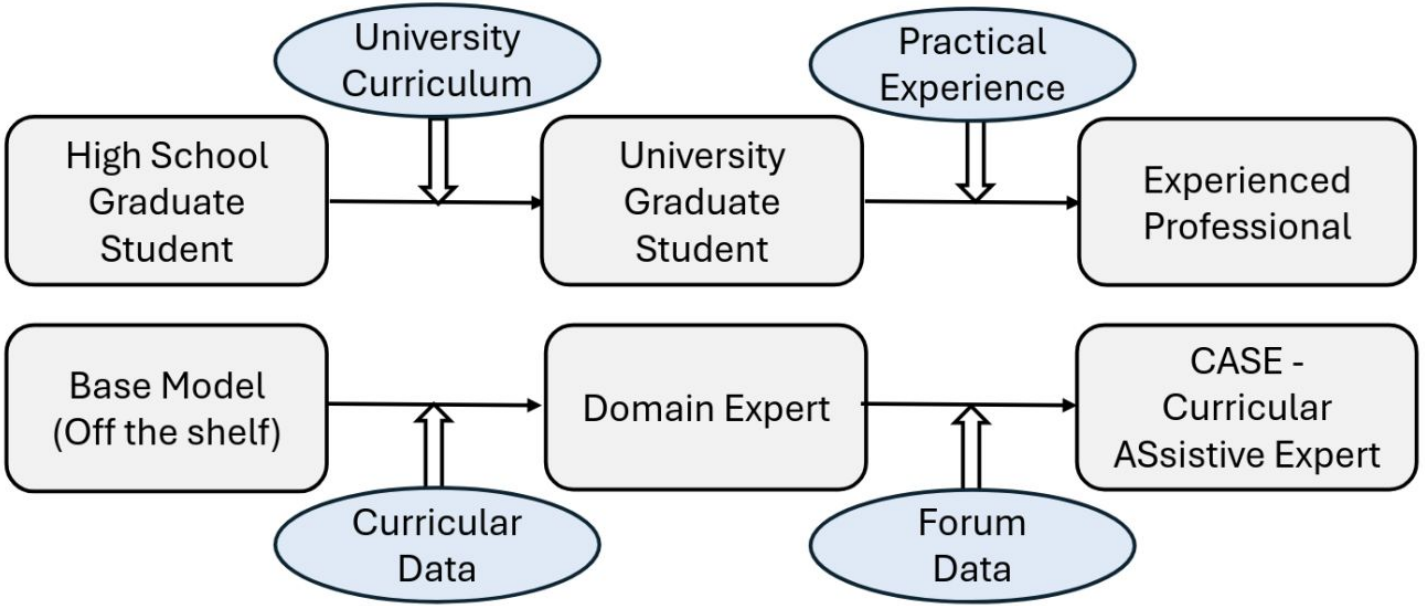
3. For a data sensitive domain like Psychology, it is not possible to get this much data.

Curricular Data to the rescue

1. “Textbooks are all you need” (Gunasekar et al) relied on ‘textbook quality data’ to pre-train a generative model.

2. We mimic the training process of a professional psychologist while further pre-training a discriminative model

Our hypothesis



The analogy between our pre-training philosophy and the steps of training required to become an experienced professional. In contrast to previous work, we ensure that relevant data is provided to the model to identify and understand the task during the pre-training stage.

Curricular Data Curation

1. We created a curricular text dataset with the help of the clinical psychologists at NIMHANS who collaborated in this work. The psychologists were a combination of people working in academia as well as professional psychiatric practitioners.

2. This dataset consisted of 110 curricular text materials that are used to train Psychology students in graduate-level education in North America, Asia, and Europe.

3. We cover the following topics like: Psychology, Counselling, Interviewing, etc

Fine Tuning Datasets and Training Setup

Fine Tuning Datasets

1. CounselChat Dataset
2. Depression_Reddit (Pirina and Çöltekin, 2018)
3. Dreaddit (Turcan and McKeown, 2019)

Further Pre-training: Nvidia P100 workbench for 60 epochs which took about 8.3 hours to train

Fine Tuning - 2x Nvidia T4 GPUs workbench for 3 epochs which took 10 mins

Performance Evaluation

We compared the performance of our model with the following available mental health models as well as general models (both generative and discriminative) -

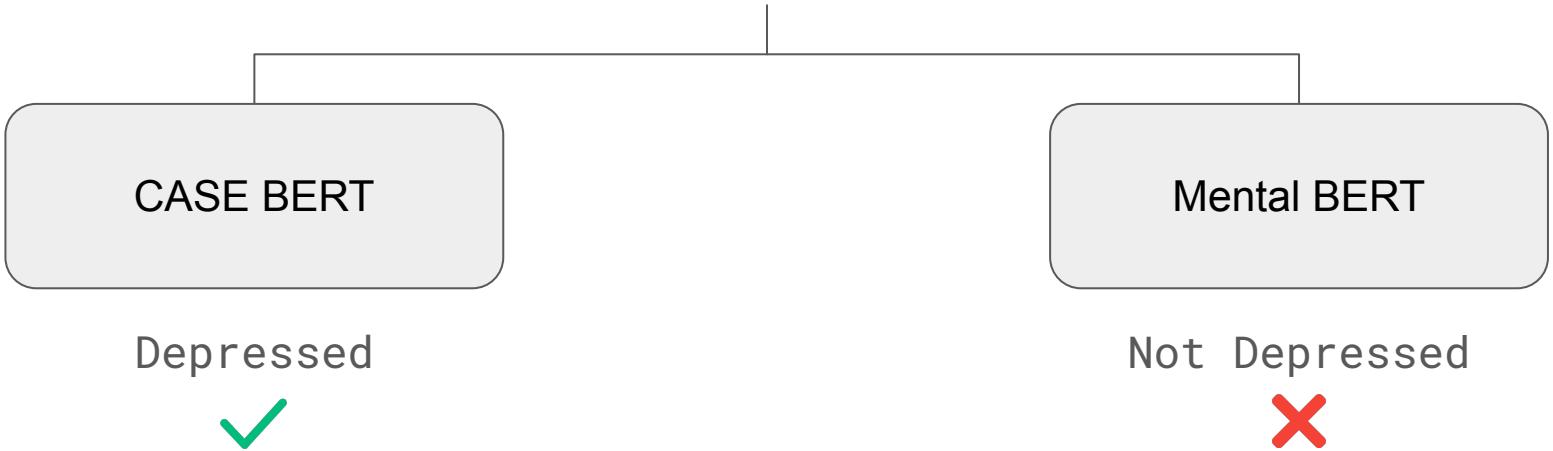
1. BERT
2. RoBERTa
3. Mental-BERT
4. Psych-Search
5. BioBERT
6. ClinicalBERT

Generative Models

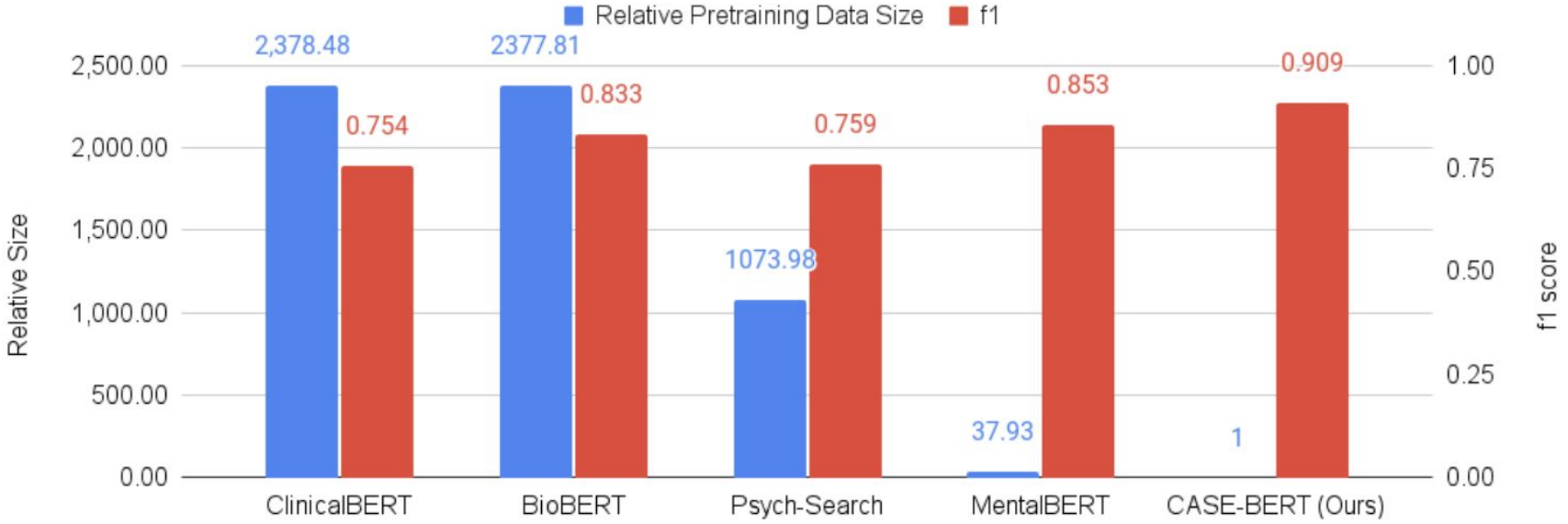
1. Mistral 7B
2. Llama 3 8B
3. Gemma 7B instruct
4. GPT-3.5-Turbo-1106

Example

“how do i get over a slump i do not know what is wrong with me at times i can be really happy excited i will talk fast and i want to do things but lately i have been staying up way later and sleeping too much…”



Data Efficiency



Comparison of our pre-training data size and our model performance on the task of identifying Depression on the CounselChat dataset against ClinicalBERT (Alsentzer et al., 2019), BioBERT (Lee et al., 2019), Psych-Search (NLP4Good, 2021) and MentalBERT (Ji et al., 2022b). We keep our dataset size as the reference and compare the size of other datasets (based on the number of words) to it

Performance Comparison

	Model Name	Depression				Anxiety			
		Recall	Precision	f1 score	Accuracy	Recall	Precision	f1 score	Accuracy
Generative	Mistral 7b	0.289	0.643	0.105	0.322	0.293	0.068	0.111	0.301
	Llama 8b Instruct	**0.974**	0.272	0.425	0.638	0.756	0.143	0.240	0.290
	Gemma 7b Instruct	0.526	0.101	0.169	0.290	0.585	0.114	0.190	0.261
	GPT-3.5-Turbo-1106	**0.974**	0.234	0.378	0.558	**0.902**	0.266	0.411	0.616
Discriminative	BERT	0.737	0.966	0.836	0.960	0.829	0.872	0.850	0.957
	RoBERTa	0.842	0.742	0.789	0.942	0.642	0.857	0.734	0.924
	Mental-BERT	0.763	0.967	0.853	0.964	0.854	0.860	0.857	0.957
	Psych-Search	0.789	0.732	0.759	0.931	0.854	0.874	0.864	0.960
	BioBERT	0.789	0.882	0.833	0.957	0.829	0.883	0.855	**0.971**
	ClinicalBERT	0.684	0.839	0.754	0.938	0.829	0.829	0.829	0.949
	CASE-BERT-Small (ours)	0.842	0.914	0.877	0.967	0.829	**0.895**	0.861	0.960
	CASE-BERT-Base (ours)	0.856	**0.969**	**0.909**	**0.975**	0.887	0.875	**0.881**	0.964

Performance Comparison using Recall, Precision, f1 score and Accuracy on the Counsel Chat Dataset (Bertagnolli, 2020) for the classes Depression and Anxiety. The best-performing numbers are highlighted in bold while the second best are underlined