Benchmarking of Transformer-Based Pre-Trained Models on Social Media Text Classification Datasets

Yuting Guo*[1], Xiangjue Doing*[1], Mohammed Ali Al-Garadi[2], Abeed Sarker[2], Cécile Paris[3], Diego Mollá-Aliod[4]

[1]Department of Computer Science, [2]Department of Biomedical Informatics, Emory University

[3]CSIRO Data61, [4]Department of Computing, Macquarie University

INTRODUCTION AND METHODOLOGY

We **compared** 3 transformer-based models (encoders), **analyzing the differences in performances** between **domain-specific** (medical), **source-specific** (social media), and **generic** pre-trained models.

- **ClinicalBioBERT (CL)** (Alsentzer et al., 2019): trained on PubMed research articles and clinical notes.
- **BERTweet (BT)** (Nguyen et al., 2020a): trained on English tweets.
- **RoBERTa-base (RT)** (Liu et al., 2019): trained on Book Corpus and English Wikipedia.

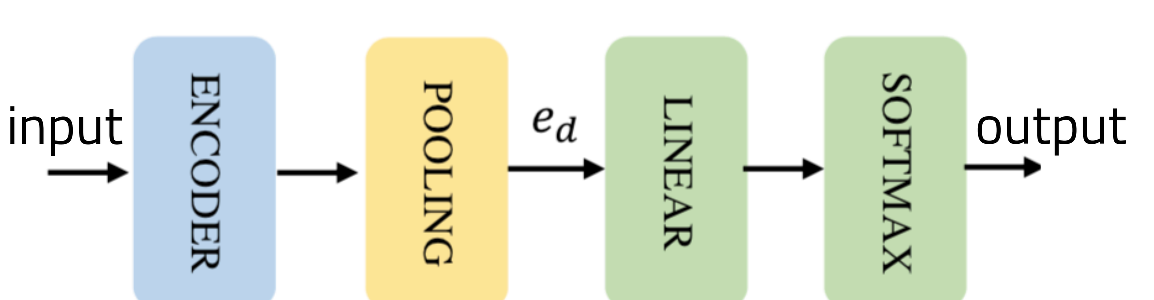

Figure 1: The framework for text classification.

- Input: training/test data
- Output:
  - Training phase: a probability vector used to compute a loss.
  - Inference phase: the class with the highest probability.

RESULTS AND DISCUSSION

Findings:

- Pre-training on relatively small source-specific data (e.g., BERTweet) may effectively benefit the downstream source-specific tasks.
- Large amount of pre-training data (e.g., RoBERTa-base) can boost the generalizability of models.
- Pre-training on small in-domain data (e.g., ClinicalBioBERT) may not benefit target tasks within the domain.

	Dataset	TRN	TST	L	S	RT	BT	CL
Health	ADR Detection	4318	1152	2	Twitter	91.4	**92.7**	90.4
	BreastCancer	3513	1204	2	Twitter	**93.9**	93.6	91.2
	PM Abuse	11829	3271	4	Twitter	81.4	**82.4**	77.4
	SMM4H-17-task1	5340	6265	2	Twitter	**93.6**	93.5	92.7
	SMM4H-17-task2	7291	5929	3	Twitter	78.4	**79.7**	75.0
	WNUT-20-task2	6238	1000	2	Twitter	**89.1**	88.3	86.5
Non-Health	OLID-1	11916	860	2	Twitter	85.1	**85.2**	83.5
	OLID-2	11916	240	2	Twitter	89.4	**90.0**	89.0
	OLID-3	11916	213	3	Twitter	69.5	**70.0**	66.4
	TRAC-1-1	11999	916	3	Facebook	58.6	**59.2**	55.4
	TRAC-1-2	11999	1257	3	Twitter	58.8	**65.8**	58.0
	TRAC-2-1	4263	1200	3	YouTube	72.8	**73.3**	63.9
	TRAC-2-2	4263	1200	2	YouTube	85.8	85.5	**87.2**
	sarcasm-1	3960	1800	2	Reddit	67.3	**69.5**	64.6
	sarcasm-2	4500	1800	2	Twitter	73.2	**76.1**	68.2
	CrowdFlower	28707	8101	13	Twitter	39.9	**41.3**	38.8
	fb-arousal-1	2085	580	9	Facebook	46.6	45.3	**46.8**
	fb-arousal-2	2088	590	9	Facebook	**54.9**	54.8	54.1
	fb-valence-1	2064	595	8	Facebook	60.2	**64.4**	54.5
	fb-valence-2	2066	604	9	Facebook	**52.8**	52.6	45.9
	SemEval-18-A	1701	1002	4	Twitter	52.3	**54.6**	46.0
	SemEval-18-F	2252	986	4	Twitter	**69.3**	67.4	65.3
	SemEval-18-J	1616	1105	4	Twitter	47.7	**51.5**	45.3
	SemEval-18-S	1533	975	4	Twitter	**54.9**	53.9	48.4
	SemEval-18-V	1182	938	8	Twitter	45.5	**46.6**	36.2

Table 1: Statistics of data sets and accuracies on the test splits. L: #classes; S: sources; **bold**: the best result; underlined: the statistically significant result compared to the next best model.

Suggestions:

- For health-related tasks on social media, it might be better to choose a source-specific pre-trained model (e.g., BERTweet for social media) rather than a domain-specific one.
- For social media text classification tasks, we recommend the use of RoBERTa-base, BERTweet or models pre-trained in similar fashion; we do not recommend the use of ClinicalBioBERT, even for health-related social media tasks.

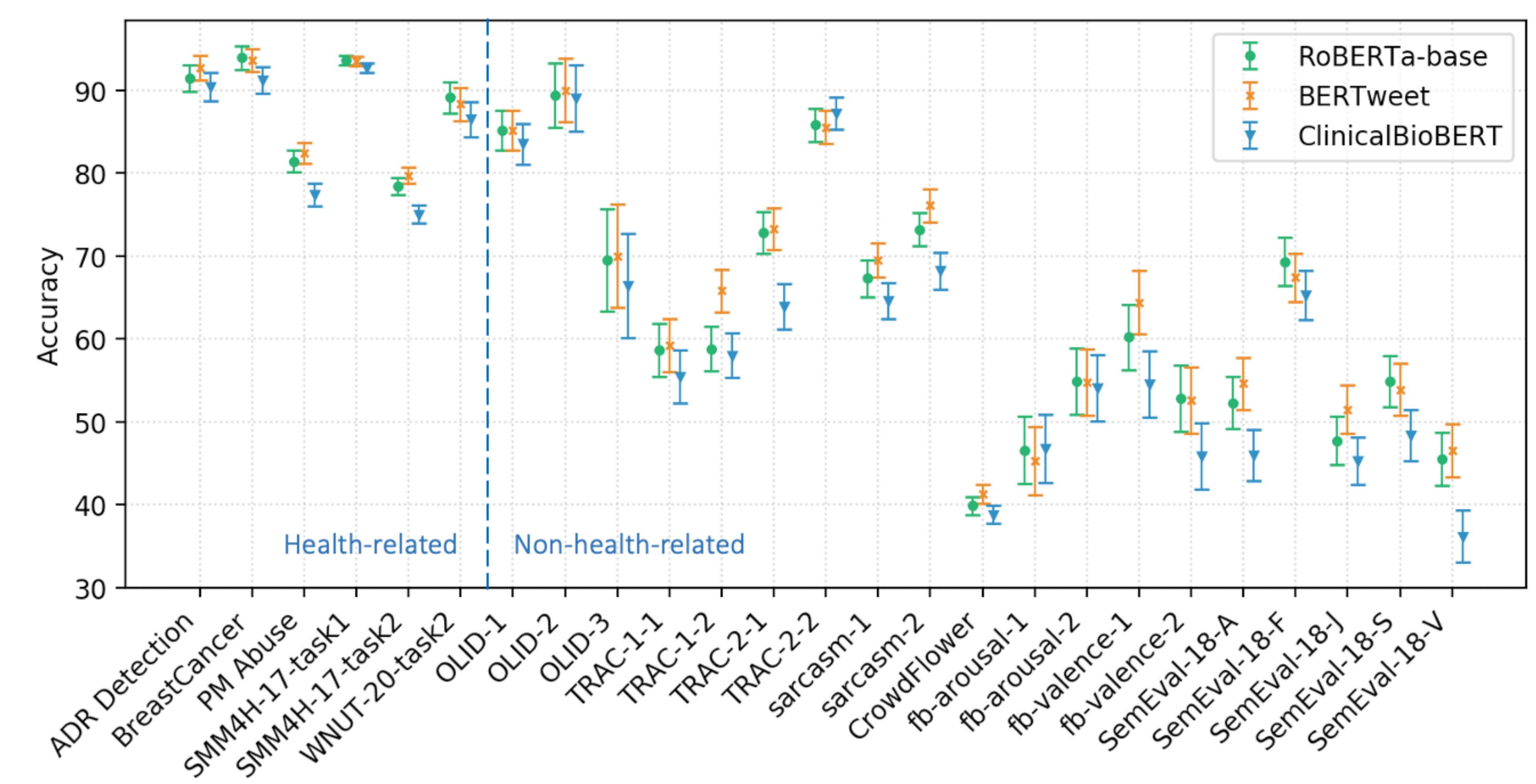

Figure 2: The 95% confidence intervals of the 3 models on our datasets.

REFERENCES

- Alsentzer, Emily, et al. "Publicly available clinical BERT embeddings." ClinicalNLP. 2019.
- Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv. 2019
- Nguyen, Dat Quoc et al. "BERTweet: A pre-trained language model for English Tweets." EMNLP. 2020.