BERTifying Sinhala - A Comprehensive Analysis of Pre-trained Language Models for Sinhala Text Classification

Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga, Sanath Jayasena

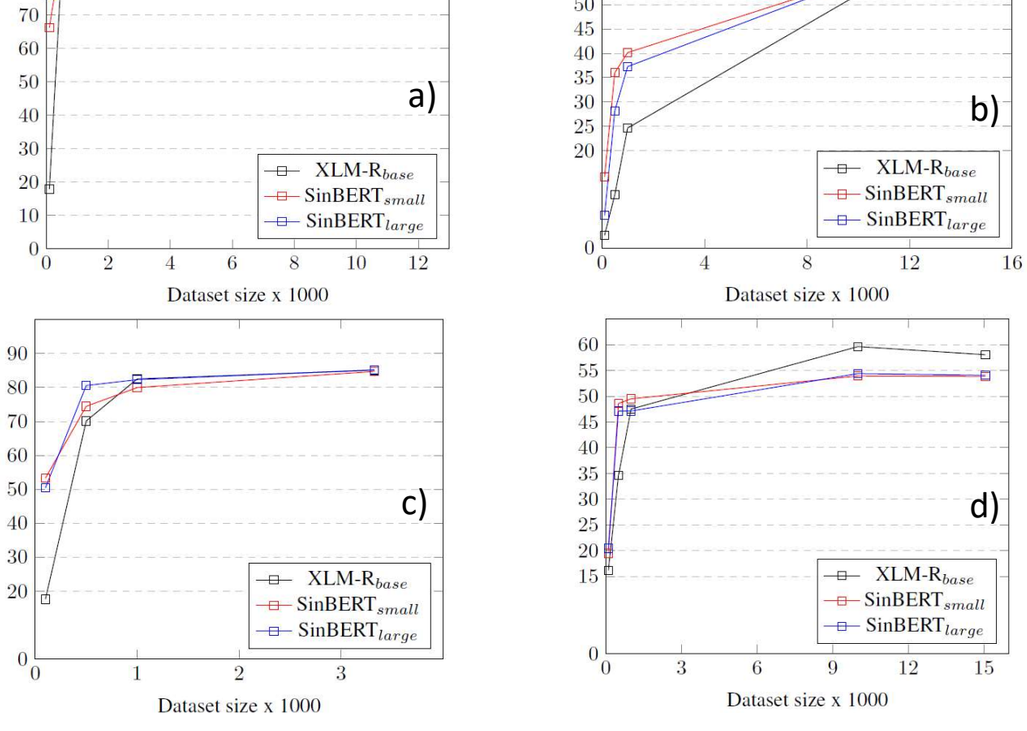
Computer Science & Engineering
University of Moratuwa

- **Sinhala;** a highly inflectional, morphologically rich language (*Indo-Aryan) ~ roughly 17M speakers*
- A resource-poor language, [*Joshi et. al 2020]*
- Multilingual language models (MLMs) could be utilized. There are monolingual language models as well. **Which is better…?**
- No conclusive finding that either type is consistent across every Task for all languages (e.g. – PhoBERT [*Nguyen et. al 2020]*/CamemBERT *[Martin et. al 2019])*

- **Identify the best models for Sinhala text classification tasks**
- **A new pre-trained monolingual model; SinBERT *(small/large)***
- Recommendations for text classification with LMs for Sinhala
- New annotated Sinhala datasets for text classification

- We experiment with,
  Multilingual: *XLM-R, LaBSE, LASER*
  Monolingual : *SinBERT, SinhalaBERTo, SinBERTo*
- 4 classification tasks,
  a) *Writing style – 4 writing genres*
  b) *News source – 9 news sources*
  c) *News category – 5 news domains*
  d) *Sentiment – 4 sentiment classes*

Model	Sentiment	News sources	News categories	Writing style
Baseline	$59.42_{w.F1}$	-	-	-
LaBSE	20.63	11.85	24.09	-
LASER	54.07	28.84	48.54	87.06
XLM-R_{base}	58.08	58.29	85.12	96.89
XLM-R_{large}	**60.45** ($68.1_{w.F1}$)	**61.84**	**89.54**	**98.41**
SinBERTo	50.83	57.22	78.07	93.84
SinhalaBERTo	49.71	57.34	82.73	94.10
$SinBERT_{small}$	53.85	60.42	84.75	95.00
$SinBERT_{large}$	54.08	60.51	85.19	95.49



https://huggingface.co/NLPC-UOM

Joshi et. al 2020 - The state and fate of linguistic diversity and inclusion in the nlp world. arXiv preprint arXiv:2004.09095
Nguyen et al 2020 - Phobert: Pre-trained language models for vietnamese. arXiv preprint arXiv:2003.00744.
Martin et al 2019 - CamemBERT: a Tasty French Language Model arXiv:1911.03894