UNSW SYDNEY

English comes in different flavours

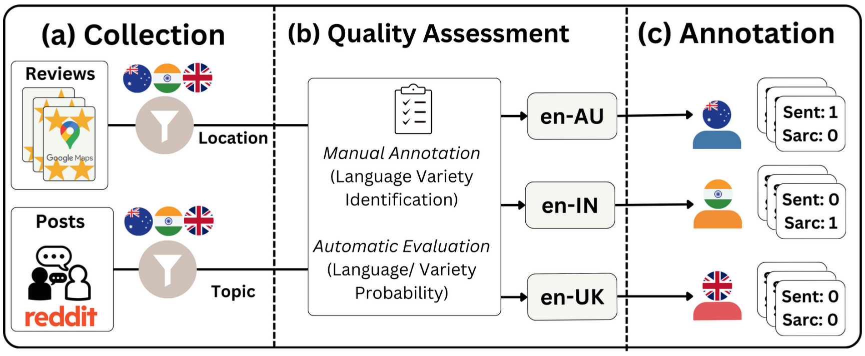
Does your LLM understand them?

UNIVERSITY OF SURREY

Try BESSTIE

A Benchmark for Sentiment and Sarcasm Classification for Varieties of English

How do we create the dataset?



Dataset Features

3 varieties of English

- + Australian (en-AU)
- + British (en-UK)

Inner circle

- + Indian (en-IN) → **Outer-circle**

Dataset with **2** annotated labels

- + Sentiment
- + Sarcasm

Dataset Statistics

Variety	Subset	Train	Valid	Test
en-AU	GOOG	946	130	270
	REDD	1763	241	501
en-IN	GOOG	1648	225	469
	REDD	1686	230	479
en-UK	GOOG	1817	248	517
	REDD	1007	138	287
Total		8867	1212	2523

We benchmark 9 LLMs

6 Encoders

- + BERT
- + DistilBERT
- + RoBERTa
- + mBERT
- + mDistilBERT
- + XLM-RoBERTa

3 Decoders

- + Gemma
- + Mistral
- + Qwen

Multilingual

Domain-Task	en-AU	en-IN	en-UK
GOOG-Sent	0.94	0.64	0.86
REDD-Sent	0.78	0.69	0.78
REDD-Sarc	0.62	0.56	0.58
Mean	0.78	0.63	0.74

Key Results for Future Research

- **Lower** model performance for **en-IN** variety.
- **Sarcasm** detection is a **harder task** than sentiment classification.
- **Error Analysis**, *in the paper*, shows the need for **focused** solutions for each **variety**.

Potential Applications

Apart from **Academic Research,** BESSTIE can be used to:

- Develop **dialogue agents** that cater to requests made by **diverse** users.
- **Localise** language models—what sounds **sarcastic** in one country might not be in another.
- Improve **social media monitoring** to prevent misinterpretation for **content moderation**.

and many more...

Srirag, D., Joshi, A., Painter, J., Kanojia, D. (2024). BESSTIE: A benchmark for sentiment and sarcasm classification for varieties of English. arXiv.

Funded by
Google Research