BESSTIE: A Benchmark for Sentiment and Sarcasm Classification for Varieties of English

Dipankar Srirag[1], **Aditya Joshi[1]**, Jordan Painter[2], Diptesh Kanojia[2]
[1]University of New South Wales, Sydney, Australia
[2]Institute for People-Centred AI, University of Surrey, Surrey, United Kingdom

What is BESSTIE?

BESSTIE is a manually labelled dataset for varieties of English evaluated on *nine* language models.

Language Varieties
- **en-AU:** Australian English
- **en-IN:** Indian English
- **en-UK:** British English

Data Sources
- **GOOGLE:** Google Places reviews
- **REDDIT:** Reddit posts and comments

Tasks (Boolean)
- Sentiment classification
- Sarcasm classification

Why would you want to use BESSTIE?

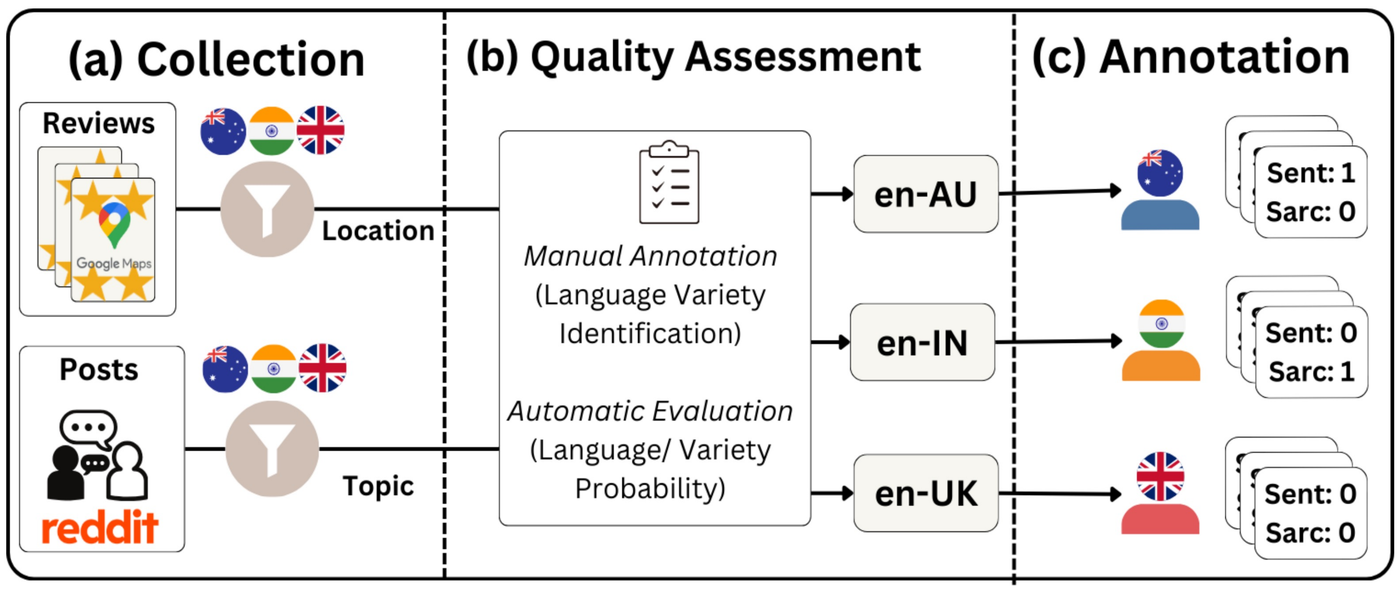
Variety	Subset	Train	Valid	Test	P(eng)	P(var)
en-AU	GOOGLE	946	130	270	0.99	0.99
	REDDIT	1763	241	501	0.98	0.95
en-IN	GOOGLE	1648	225	469	0.99	0.94
	REDDIT	1686	230	479	0.87	0.78
en-UK	GOOGLE	1817	248	517	0.99	0.99
	REDDIT	1007	138	287	0.98	0.93
Total		8867	1212	2523		

Domain-Task	en-AU	en-IN	en-UK
GOOGLE-Sentiment	0.94	0.64	0.86
REDDIT-Sentiment	0.78	0.69	0.78
REDDIT-Sarcasm	0.62	0.56	0.58
Average	0.78	0.63	0.74

Dataset Creation

Location-based filtering: GOOGLE

Topic-based filtering: REDDIT



Native speakers manually annotate every instance with sentiment and sarcasm.

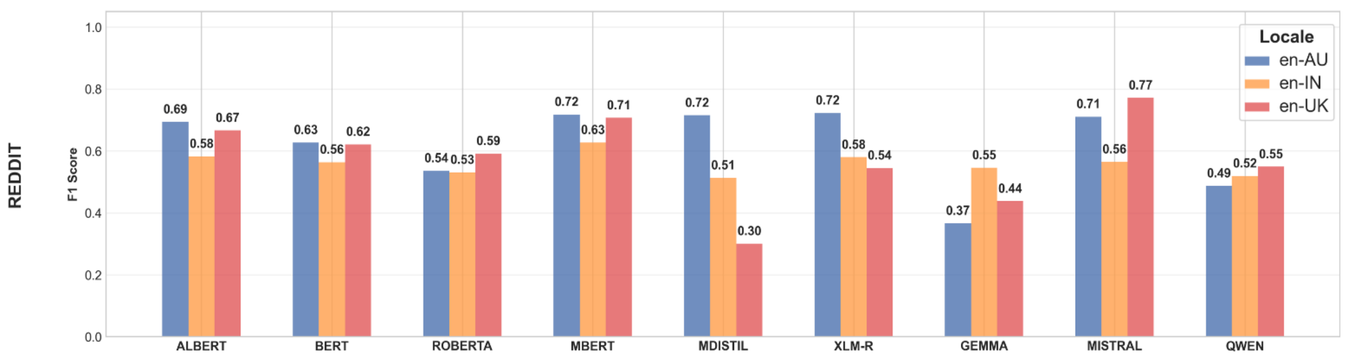
Cohen's Kappa:

Variety	Sent.	Sarc.
en-AU	0.61	0.47
en-IN	0.65	0.51
en-UK	0.79	0.63

Language Variety Probability:
Manual Annotation: Performed by Native speakers of en-IN and en-UK.
Automatic Evaluation: Using DistilBERT [2] fine-tuned using ICE-Corpus [2] (India and Australia)
Language Probability: fastText for English

Key Results

What results do we report?



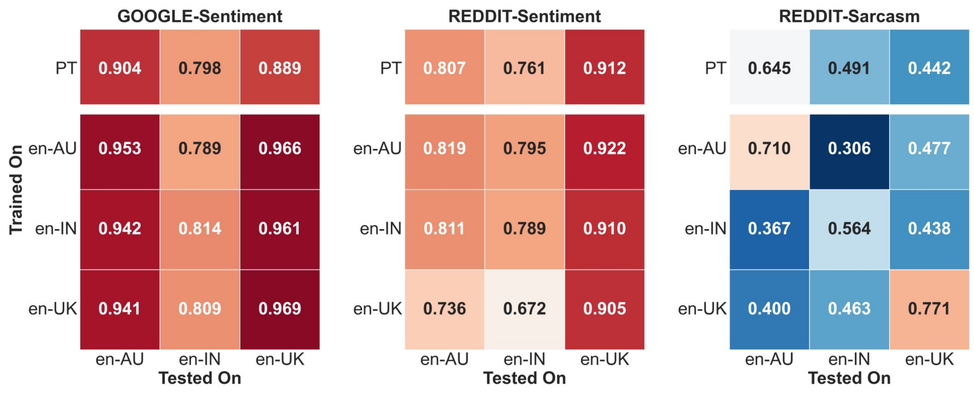
Sarcasm Classification

*Graphs for all task-model combinations in the paper

How do the models compare?

Model Properties	Average
Encoder or decoder models?	
Encoder-only	0.74
Decoder-only	0.67
Monolingual or Multilingual models?	
Monolingual	0.72
Multilingual	0.71

Can't we just train the models on any other language variety?



Cross-variety performance analysis of MISTRAL. The figure compares three different scenarios: pre- trained (PT), in-variety fine-tuning, and cross-variety fine-tuning for sentiment and sarcasm classification across all varieties.

*Cross-Domain Performance in the paper

What kind of errors do we encounter?

Variety	#samples	Dialect Features	Colloquial Expressions	Contextual Understanding	Code-mixing
en-AU	70	9	28	6	-
en-IN	90	97	33	3	8
en-UK	53	7	15	4	-

*Error Examples in the paper

ACL 2025 VIENNA JULY 27 - AUGUST 1
UNSW SYDNEY
UNIVERSITY OF SURREY
Funded by Google Research