



BESSTIE

In ACL 2025 (Findings)

A Benchmark for Sentiment and Sarcasm Classification for Varieties of English

Dipankar Srirag¹, Aditya Joshi¹, Jordan Painter², Diptesh Kanojia²

¹University of New South Wales, Sydney, Australia

²Institute for People-Centred AI, University of Surrey, Surrey, United Kingdom



UNSW
SYDNEY



UNIVERSITY OF
SURREY

Funded by
Google Research

Existing benchmarks

Benchmark	Sent.	Sarc.	Eng.	Var.
Cieliebak et al. (2017)	✓	×	×	×
Wang et al. (2018)	✓	×	✓	×
Alharbi et al. (2020)	✓	×	×	✓
Abu Farha et al. (2021)	✓	✓	×	✓
Elmadany et al. (2023)	✓	✓	×	✓
Faisal et al. (2024)	✓	×	×	✓

SST [1]

[1] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. [Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

BESSTIE

Benchmark	Sent.	Sarc.	Eng.	Var.
Cieliebak et al. (2017)	✓	×	×	×
Wang et al. (2018)	✓	×	✓	×
Alharbi et al. (2020)	✓	×	×	✓
Abu Farha et al. (2021)	✓	✓	×	✓
Elmadany et al. (2023)	✓	✓	×	✓
Faisal et al. (2024)	✓	×	×	✓
BESSTIE	✓	✓	✓	✓

BESSTIE

A benchmark for sentiment and sarcasm classification for varieties of English

Why use BESSTIE?

Why use BESSTIE?

- Standard English
 - that's far too tragic to merit such superficial treatment

Why use BESSTIE?

- Standard English
 - that's far too tragic to merit such superficial treatment
- Text with **dialect features**
 - quantity is very very less

Why use BESSTIE?

- Standard English
 - that's far too tragic to merit such superficial treatment
- Text with **dialect features**
 - quantity is very very less
- Text with **colloquial terms**
 - Whitlam wanted to shut it down did he not? **Nek Minit**
 - This case seriously is now sounding like a badly written **Crime Patrol** episode.

Dataset Creation



Dataset Creation



Language Varieties

- Australian (en-AU)
- British (en-UK)
- Indian (en-IN)

Dataset Creation



Language Varieties [2]

- | | |
|----------------------|--------------|
| • Australian (en-AU) | Inner-circle |
| • British (en-UK) | Inner-circle |
| • Indian (en-IN) | Outer-circle |

[2] Braj B. Kachru. 1990. World englishes and applied linguistics. *World Englishes*, 9(1):3–20.

Dataset Creation



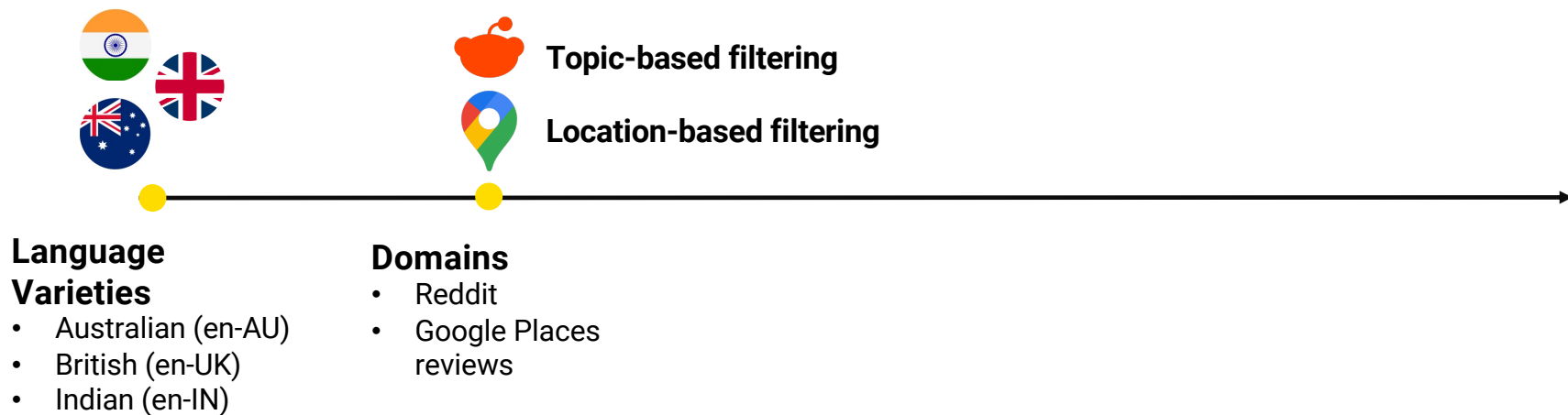
Language Varieties

- Australian (en-AU)
- British (en-UK)
- Indian (en-IN)

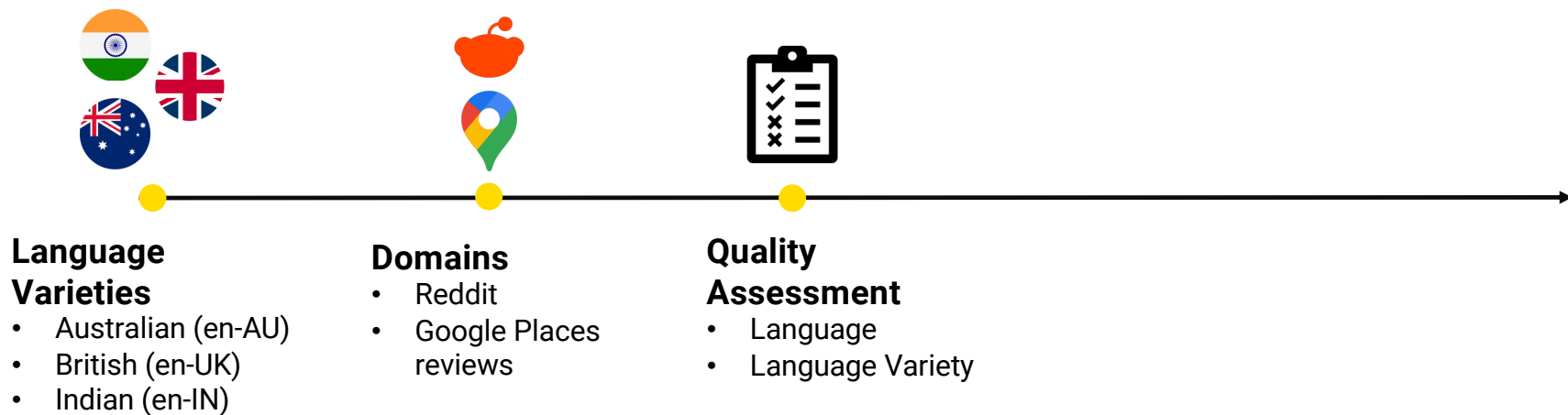
Domains

- Reddit
- Google Places reviews

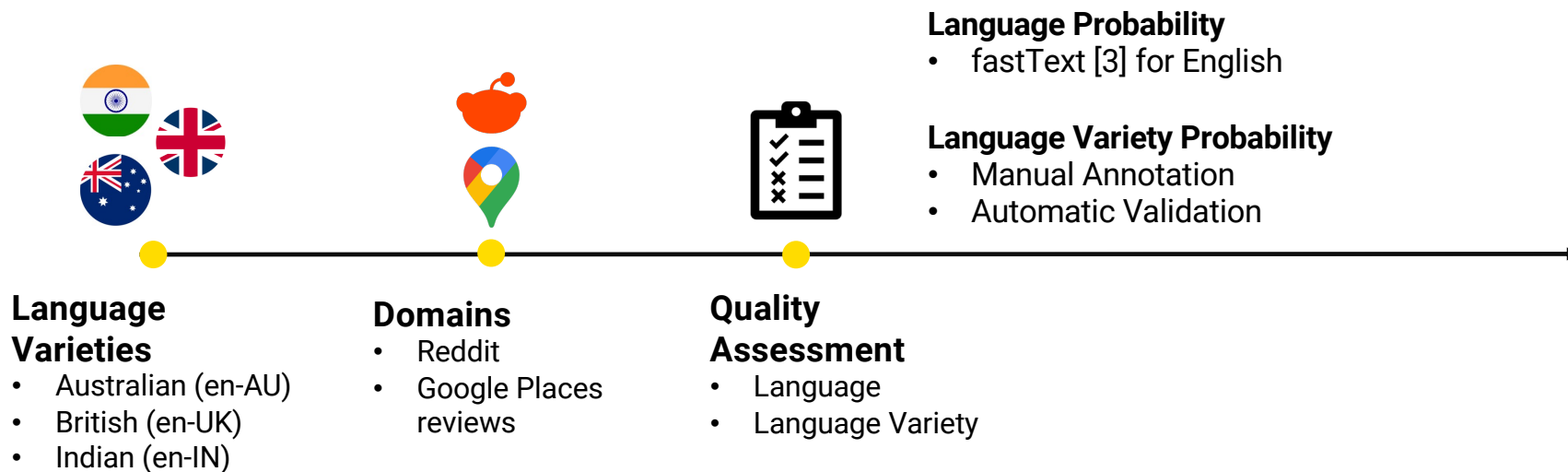
Dataset Creation



Dataset Creation



Dataset Creation



[3] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.



Quality Assessment- Language Variety

Manual Annotation

Automatic Validation



Quality Assessment- Language Variety

Manual Annotation

ant-IN	Cannot say	16	23	3	44
	en-AU	2	28	4	36
	en-IN	6	8	46	26
	en-UK	1	17	1	39
		ant-UK			

Automatic Validation

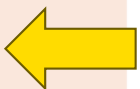


Quality Assessment- Language Variety

Manual Annotation

ant-IN	Cannot say	16	23	3	44
	en-AU	2	28	4	36
	en-IN	6	8	46	
	en-UK	1	17	1	39
		Cannot say	en-AU	en-IN	en-UK
		ant-UK			

Automatic Validation





Quality Assessment- Language Variety

Manual Annotation

Cannot say	16	23	3	44	
en-AU	2	28	4	36	←
en-IN	6	8	46	26	
en-UK	1	17	1	39	←
	Cannot say	en-AU	en-IN	en-UK	

ant-IN

ant-AU

ant-UK

Automatic Validation



Quality Assessment- Language Variety

Manual Annotation

Cannot say	16	23	3	44
en-AU	2	28	4	36
en-IN	6	8	46	26
en-UK	1	17	1	39
	Cannot say	en-AU	en-IN	en-UK

ant-UK

Two yellow arrows point from the 'en-AU' and 'en-UK' rows of the manual annotation table towards the automatic validation section.

Automatic Validation

DistilBERT [4]

Outer Circle

vs.

Inner Circle

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. Preprint, arXiv:1910.01108.



Quality Assessment- Language Variety

Manual Annotation

Cannot say	16	23	3	44
en-AU	2	28	4	36
en-IN	6	8	46	26
en-UK	1	17	1	39
	Cannot say	en-AU	en-IN	en-UK

ant-UK

Two yellow arrows point from the 'en-AU' and 'en-UK' rows of the table towards the 'Automatic Validation' section.

Automatic Validation

DistilBERT [4]

ICE-India [5]

vs.

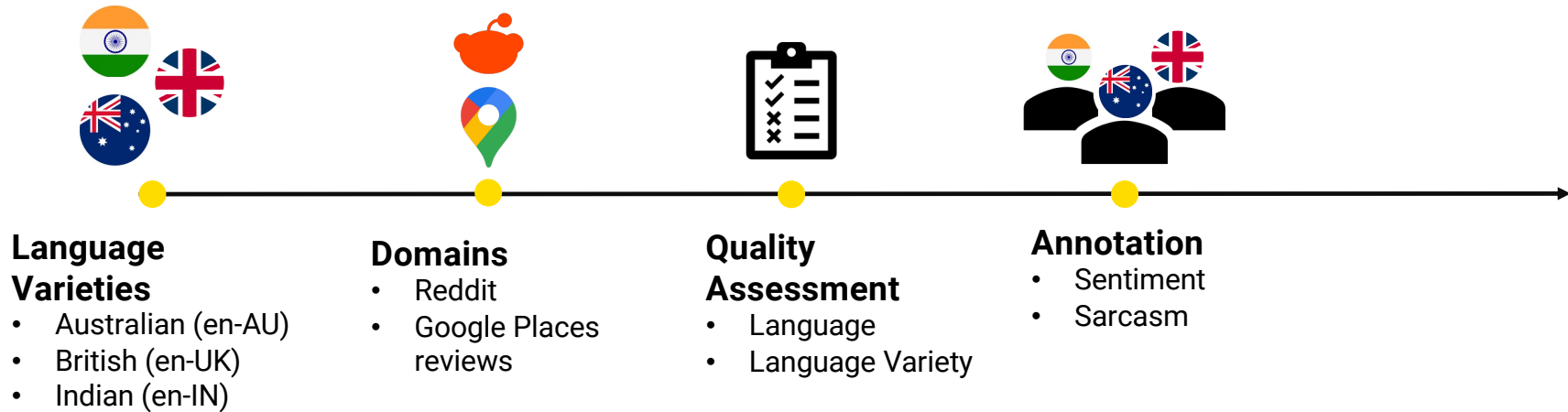
ICE-Australia
[6]

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. Preprint, arXiv:1910.01108.

[5] Sidney Greenbaum and Gerald Nelson. 1996. The international corpus of English (ICE) project. World Englishes, 15(1):3–15.

[6] Adam Smith and Pam Peters. 2023. International Corpus of English (ICE).

Dataset Creation

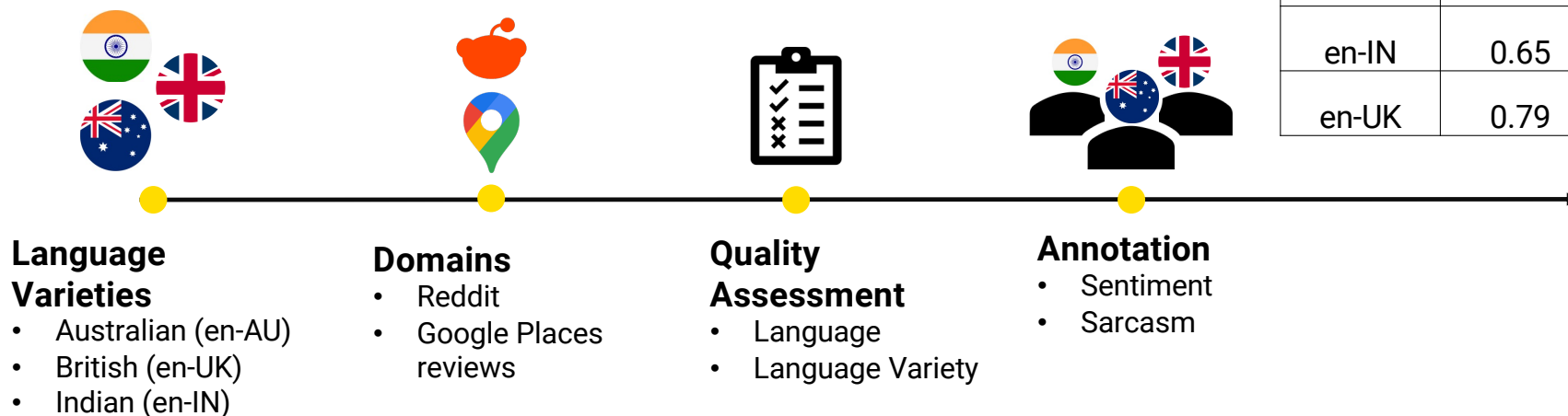


Dataset Creation

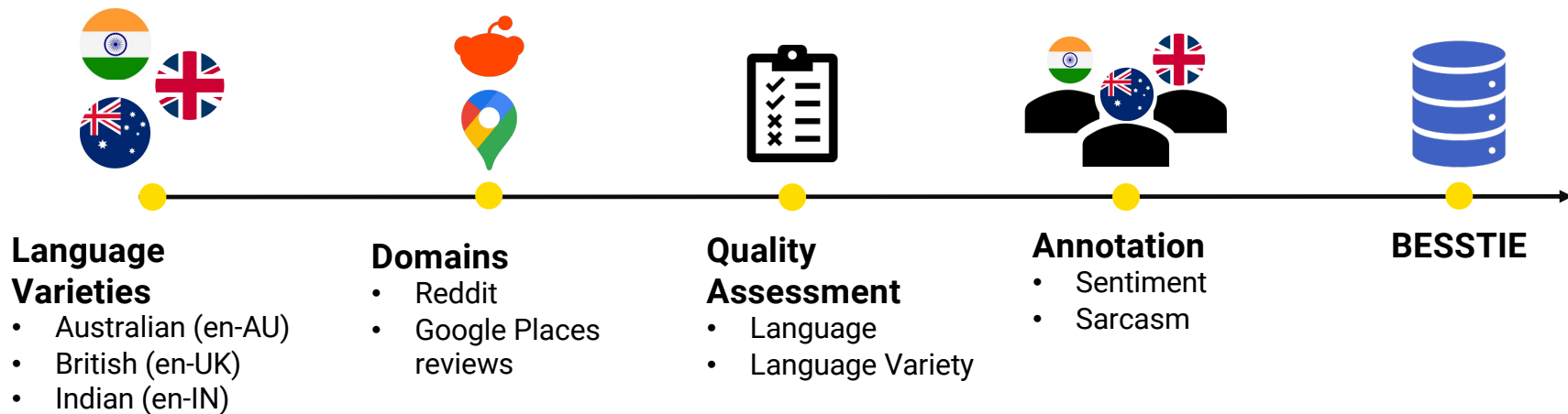
Native speakers manually annotate every instance with sentiment and sarcasm.

Cohen's Kappa:

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



Dataset Creation



Dataset Statistics

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		

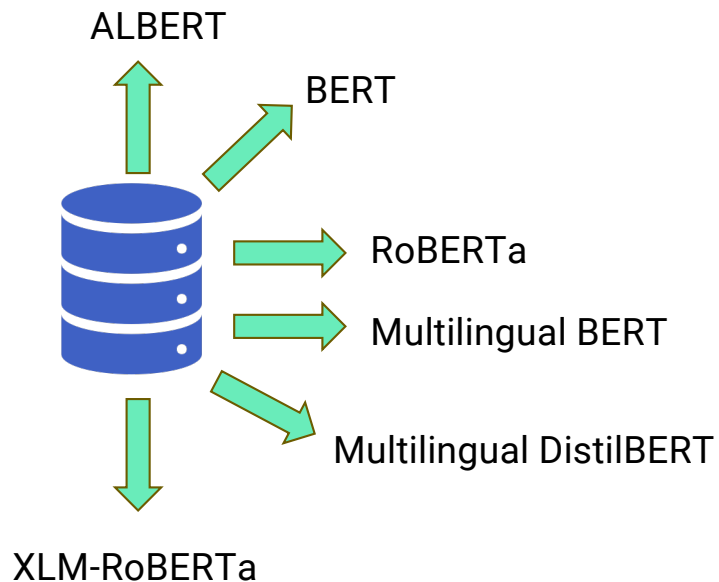
Dataset Statistics

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	415	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		

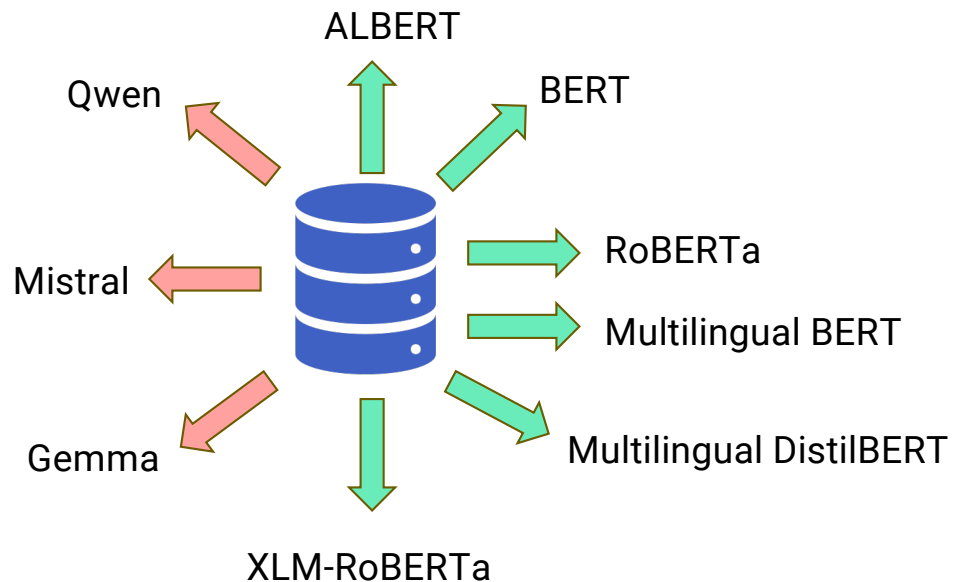
Experiments



Experiments



Experiments



Overall Results

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

Overall Results

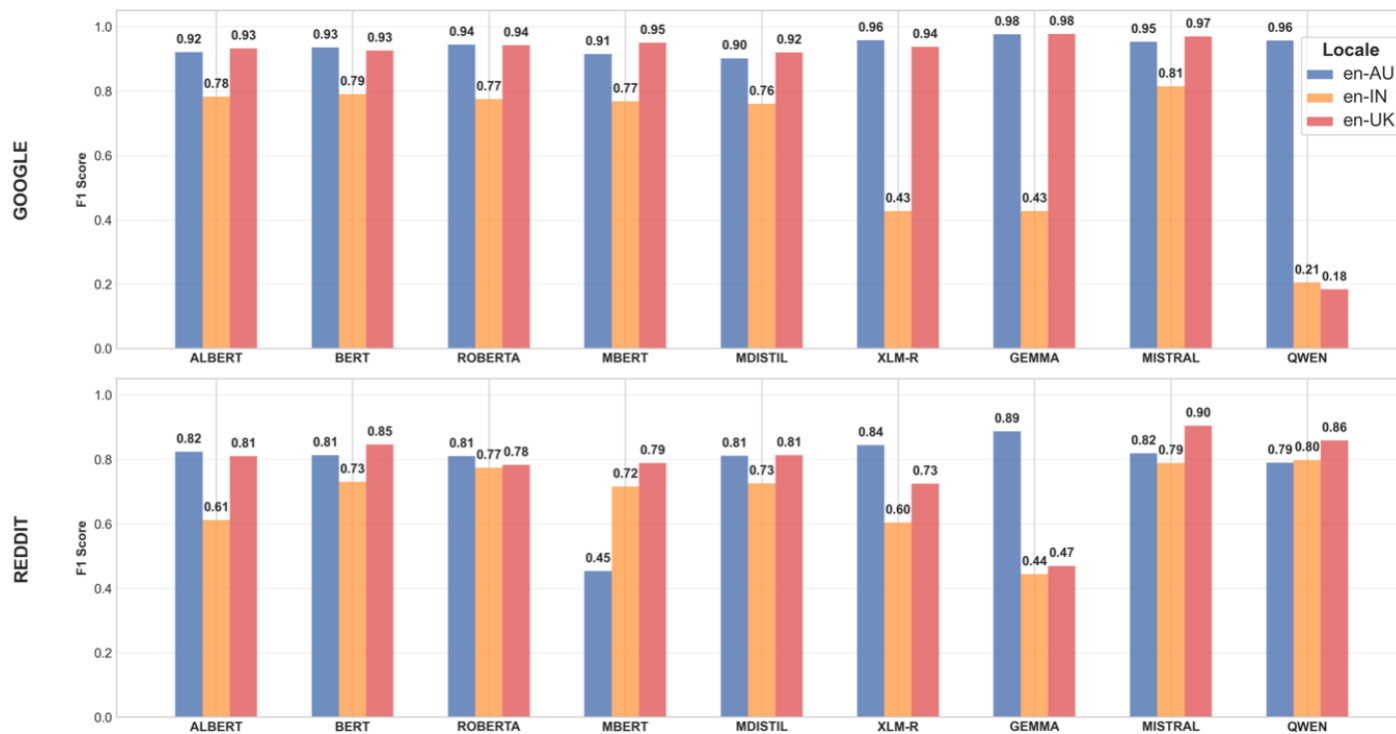
Domain-Task	en-AU	en-IN	en-UK
GOOGLE-Sentiment	0.94	0.64	0.86
REDDIT-Sentiment	0.78	0.69	0.78
REDDIT-Sarcasm	0.62	0.56	0.58

Overall Results

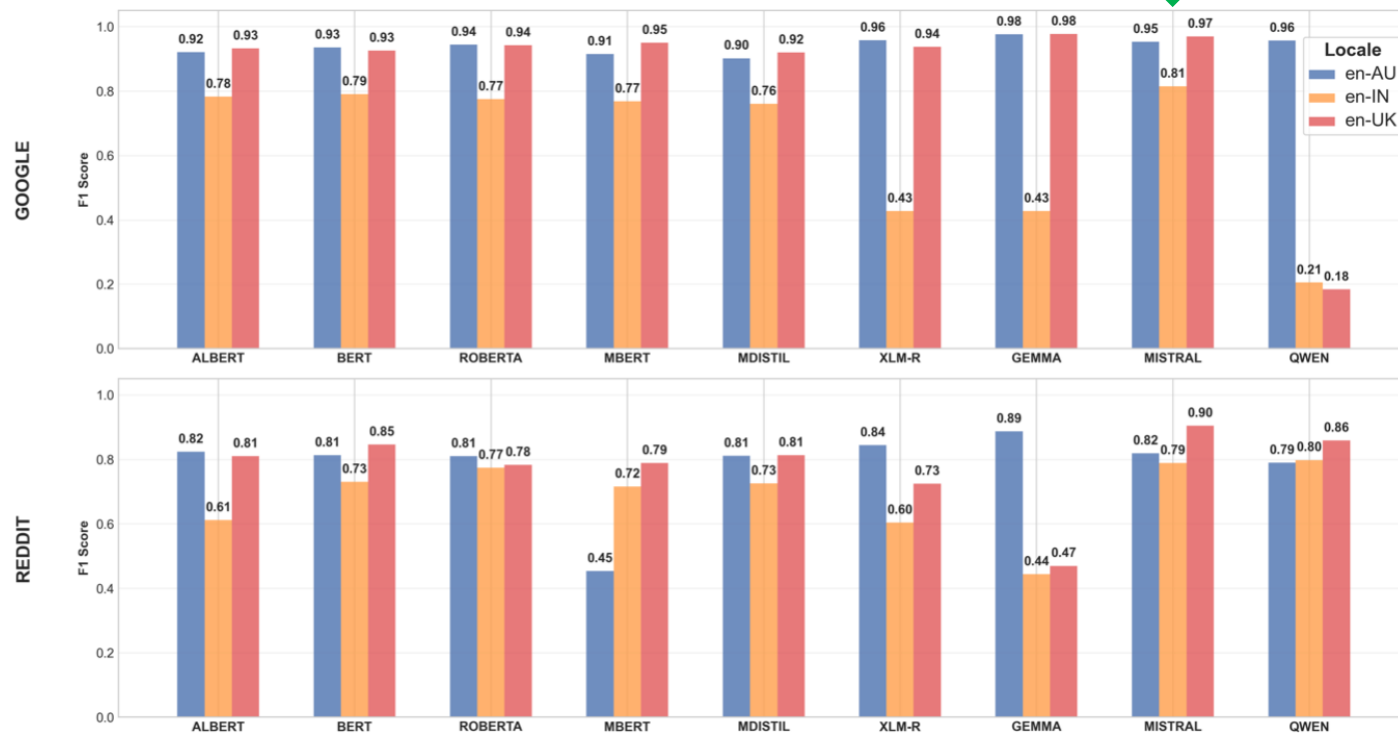
Domain-Task	en-AU	en-IN	en-UK
GOOGLE-Sentiment	0.94	0.64	0.86
REDDIT-Sentiment	0.78	0.69	0.78
REDDIT-Sarcasm	0.62	0.56	0.58



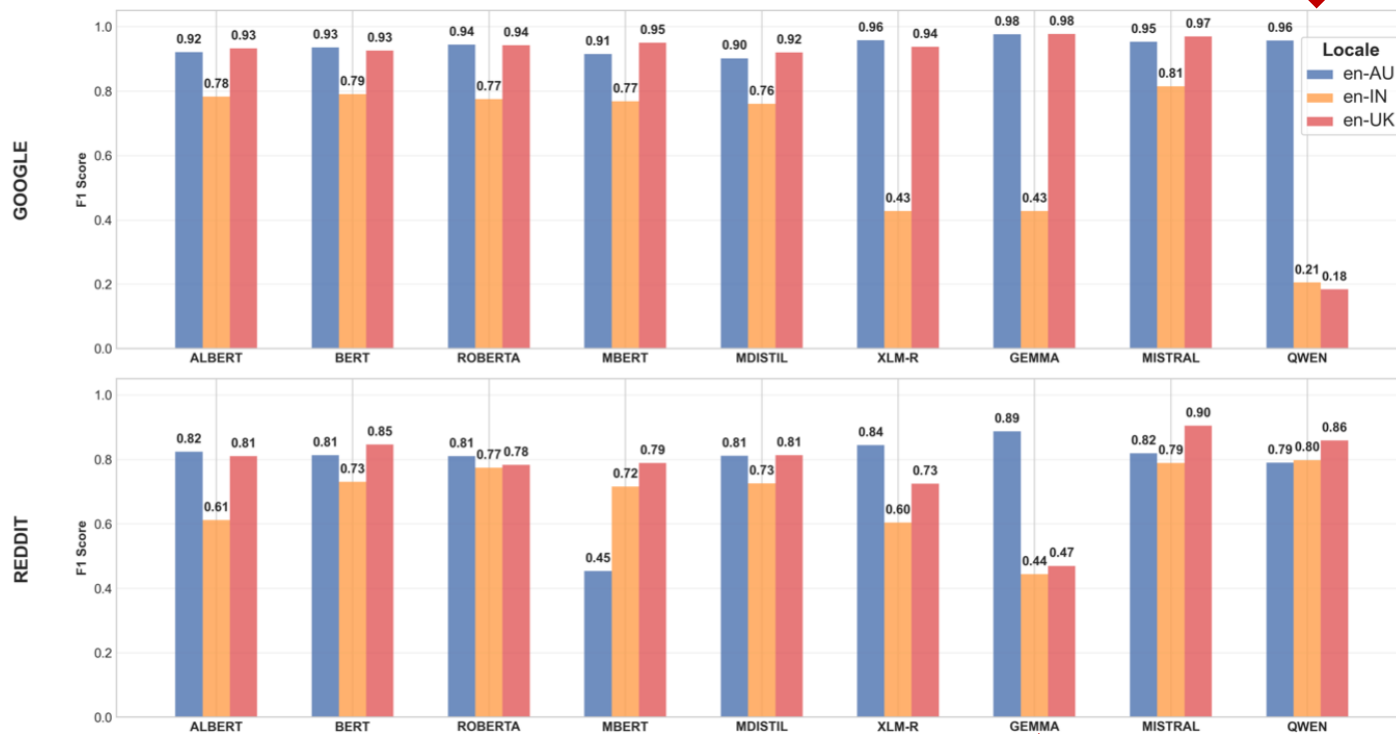
Sentiment Classification



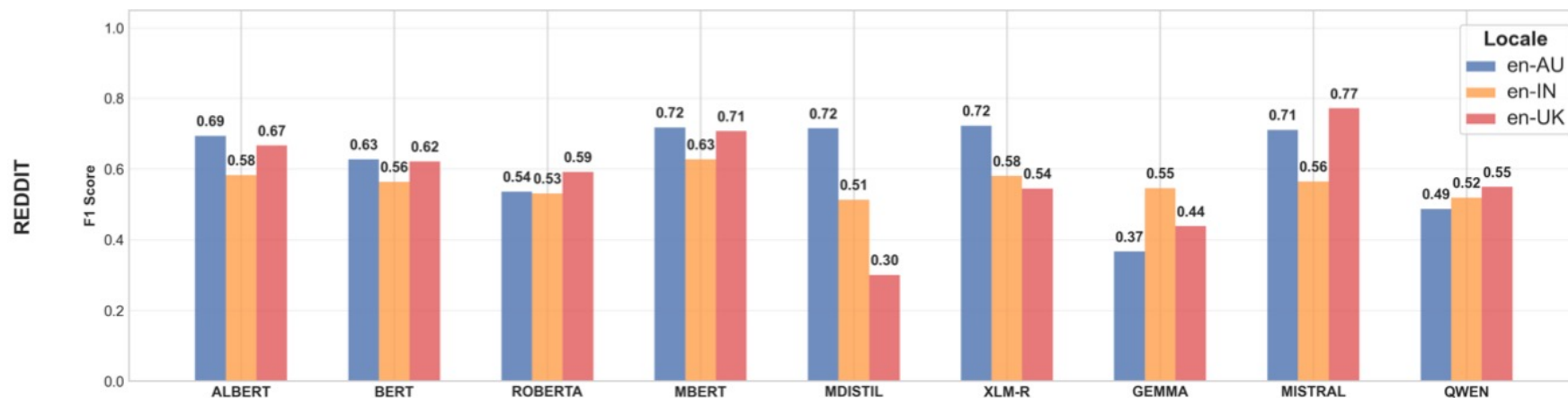
Sentiment Classification



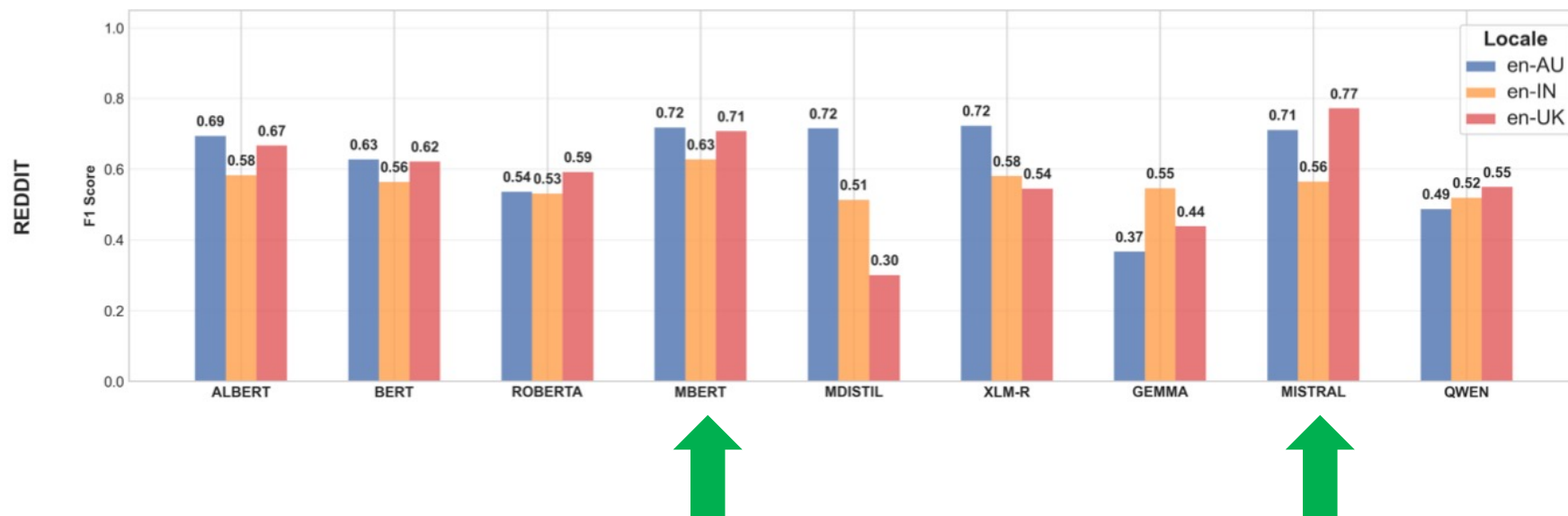
Sentiment Classification



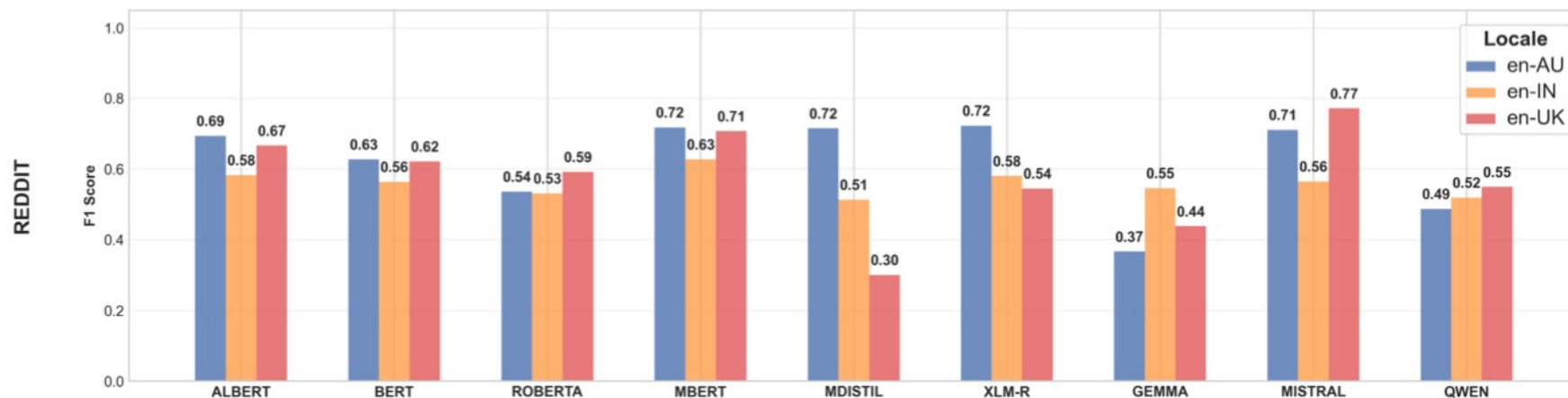
Sarcasm Classification



Sarcasm Classification



Sarcasm Classification



Results – Model Properties

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

Results – Model Properties

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

Results – Model Properties

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

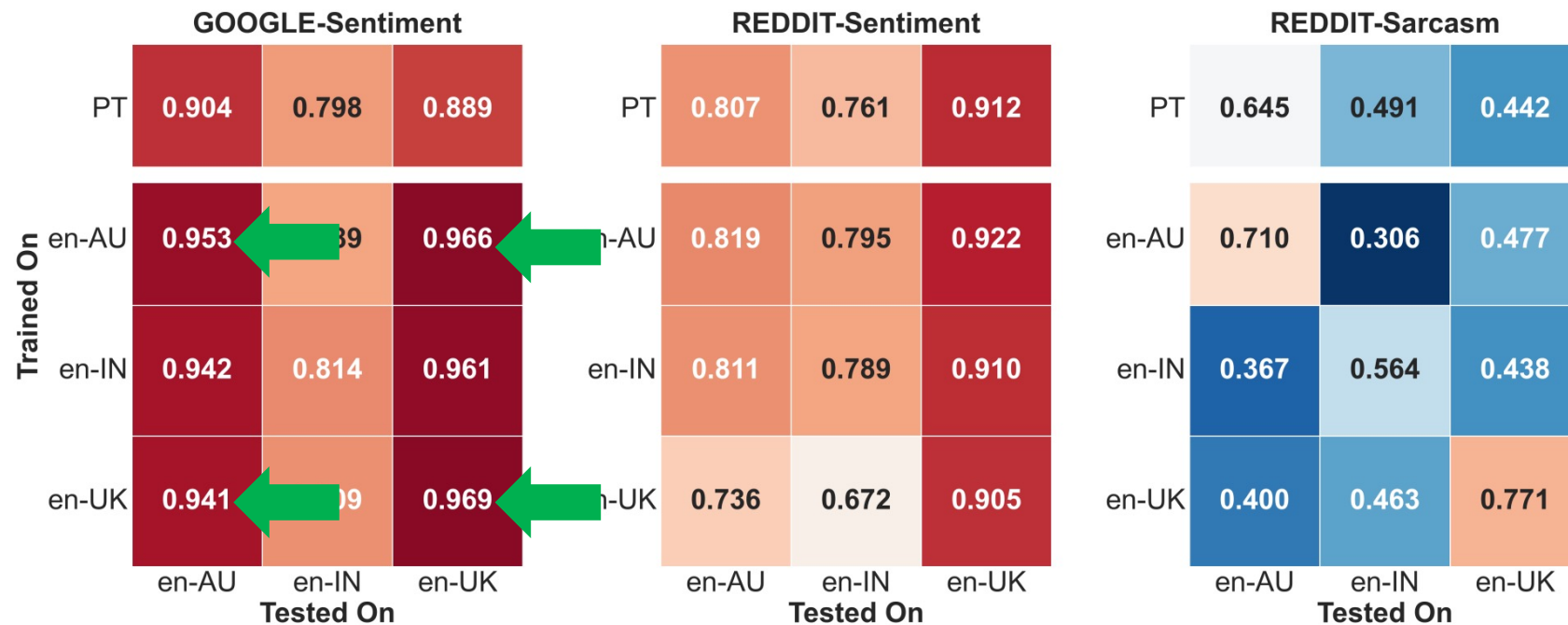
Results – Cross Variety

		GOOGLE-Sentiment		
Trained On	PT	0.904	0.798	0.889
	en-AU	0.953	0.789	0.966
	en-IN	0.942	0.814	0.961
	en-UK	0.941	0.809	0.969
		en-AU	en-IN	en-UK
		Tested On		

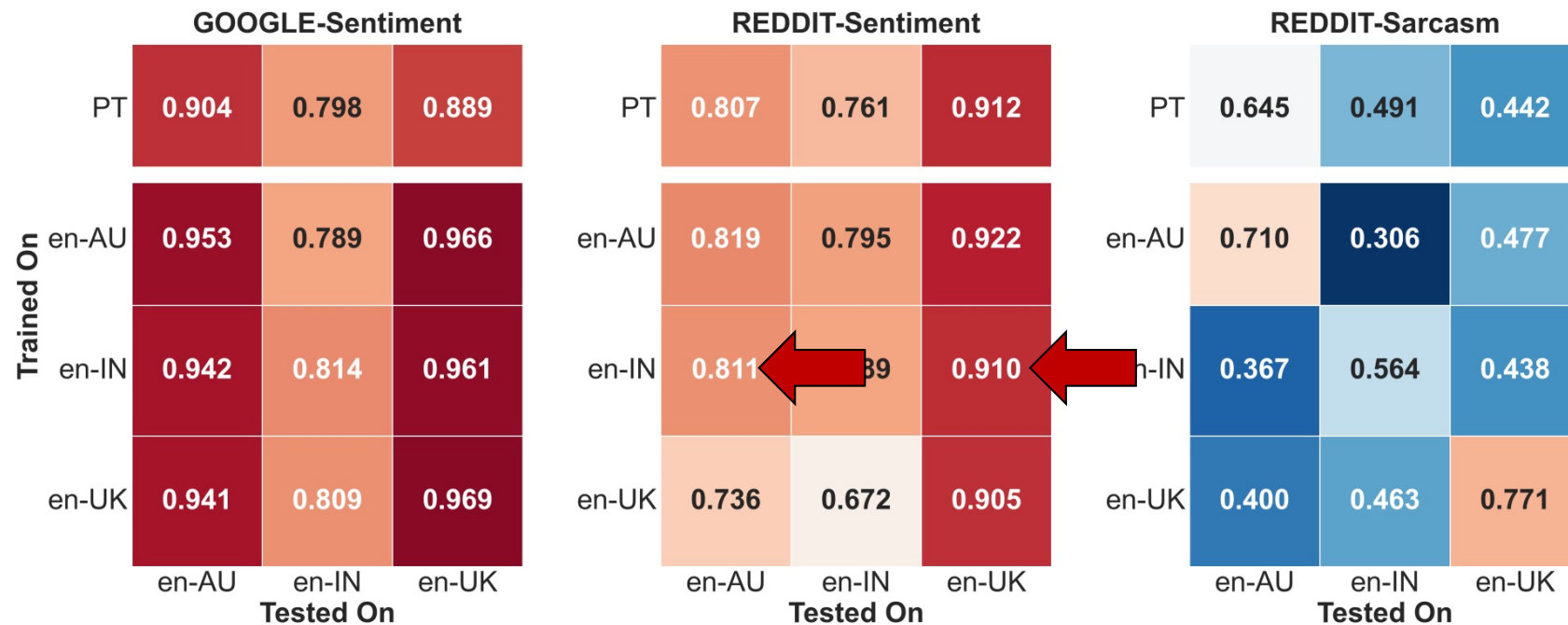
		REDDIT-Sentiment		
	PT	0.807	0.761	0.912
	en-AU	0.819	0.795	0.922
	en-IN	0.811	0.789	0.910
	en-UK	0.736	0.672	0.905
		en-AU	en-IN	en-UK
		Tested On		

		REDDIT-Sarcasm		
	PT	0.645	0.491	0.442
	en-AU	0.710	0.306	0.477
	en-IN	0.367	0.564	0.438
	en-UK	0.400	0.463	0.771
		en-AU	en-IN	en-UK
		Tested On		

Results – Cross Variety



Results – Cross Variety



Results – Cross Variety

		GOOGLE-Sentiment		
Trained On	PT	0.904	0.798	0.889
	en-AU	0.953	0.789	0.966
	en-IN	0.942	0.814	0.961
	en-UK	0.941	0.809	0.969
		en-AU	en-IN	en-UK
		Tested On		

		REDDIT-Sentiment		
Trained On	PT	0.807	0.761	0.912
	en-AU	0.819	0.795	0.922
	en-IN	0.811	0.789	0.910
	en-UK	0.736	0.672	0.905
		en-AU	en-IN	en-UK
		Tested On		

		REDDIT-Sarcasm		
Trained On	PT	0.645	0.491	0.442
	en-AU	0.710	0.6	0.477
	en-IN	0.367	0.564	0.8
	en-UK	0.400	0.463	0.771
		en-AU	en-IN	en-UK
		Tested On		

Results – Cross Variety

		GOOGLE-Sentiment		
Trained On	PT	0.904	0.798	0.889
	en-AU	0.953	0.789	0.966
	en-IN	0.942	0.814	0.961
	en-UK	0.941	0.809	0.969
		en-AU	en-IN	en-UK
		Tested On		

		REDDIT-Sentiment		
Trained On	PT	0.807	0.761	0.912
	en-AU	0.819	0.795	0.922
	en-IN	0.811	0.789	0.910
	en-UK	0.736	0.672	0.905
		en-AU	en-IN	en-UK
		Tested On		

		REDDIT-Sarcasm		
Trained On	PT	0.645	0.491	0.442
	en-AU	0.710	0.306	0.477
	en-IN	0.367	0.564	0.438
	en-UK	0.400	0.463	0.771
		en-AU	en-IN	en-UK
		Tested On		

Error Analysis

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 Analysis

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		15	4	-

quantity is very very less

Error Analysis

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	

Bridge chori ho jaaega

Error Analysis

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	-

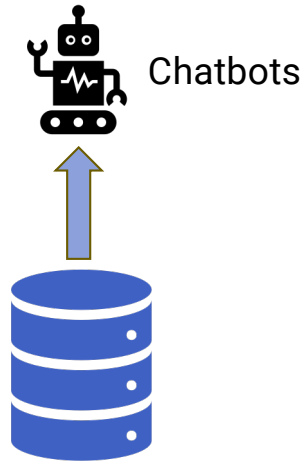


Whitlam wanted to shut it down did he not? ***Nek Minit***

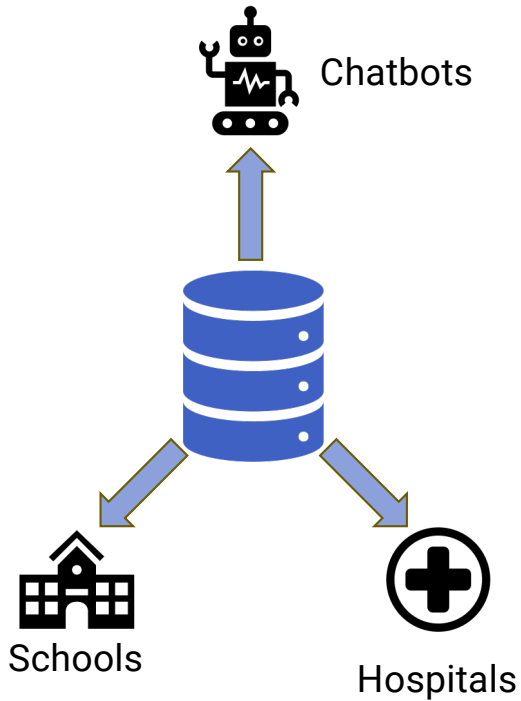
Who can use this?



Who can use this?

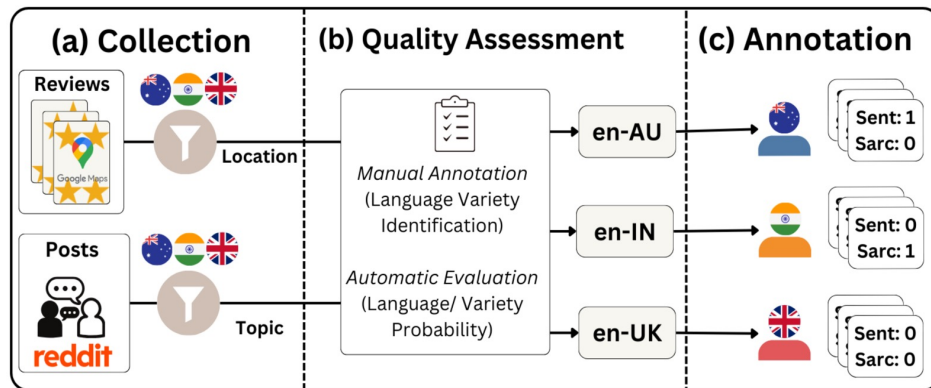


Who can use this?



BESSTIE

A Benchmark for Sentiment and Sarcasm Classification for Varieties of English



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



Find Aditya at



UNSW
SYDNEY



UNIVERSITY OF
SURREY

Funded by
Google Research