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-Centered AI, University of Surrey, Surrey, United Kingdom
{d.srirag,aditya.joshi}@unsw.edu.au {jp1106,d.kanojia}@surrey.ac.uk

Abstract

Despite large language models (LLMs) being known to exhibit bias against non-standard language varieties, there are no known labelled datasets for sentiment analysis of English. To address this gap, we introduce BESSTIE, a benchmark for sentiment and sarcasm classification for three varieties of English: Australian (en-AU), Indian (en-IN), and British (en-UK). We collect datasets for these language varieties using two methods: location-based for Google Places reviews, and topic-based filtering for Reddit comments. To assess whether the dataset accurately represents these varieties, we conduct two validation steps: (a) manual annotation of language varieties and (b) automatic language variety prediction. Native speakers of the language varieties manually annotate the datasets with sentiment and sarcasm labels. We perform an additional annotation exercise to validate the reliance of the annotated labels. Subsequently, we fine-tune nine large language models (LLMs) (representing a range of encoder/decoder and mono/multilingual models) on these datasets, and evaluate their performance on the two tasks. Our results show that the models consistently perform better on inner-circle varieties (*i.e.*, en-AU and en-UK), in comparison with en-IN, particularly for sarcasm classification. We also report challenges in cross-variety generalisation, highlighting the need for language variety-specific datasets such as ours. BESSTIE promises to be a useful evaluative benchmark for future research in equitable LLMs, specifically in terms of language varieties. The BESSTIE dataset is publicly available at: https://huggingface.co/datasets/unswnlporg/BESSTIE.

1 Introduction

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	✓	✓	✓	✓

Table 1: Comparison of BESSTIE with past benchmarks for sentiment or sarcasm classification. 'Sent.' indicates sentiment classification, 'Sarc.' denotes sarcasm classification, 'Eng.' denotes English, and 'Var.' denotes language varieties. A checkmark (✓) denotes the availability of a particular feature, while a cross (✗) indicates its absence.

Benchmark-based evaluation (Socher et al., 2013) of large language models (LLMs) is the prevalent norm in natural language processing (NLP). Benchmarks provide labelled datasets (Wang et al., 2019; Muennighoff et al., 2023) for specific tasks so that LLMs can be evaluated against them. However, most NLP benchmarks do not contain text in language varieties such as national varieties, dialects, sociolects or creoles (Plank, 2022; Lent et al., 2024; Joshi et al., 2025). As a result, despite the superlative performance of LLMs on Standard American English, LLMs are not evaluated on other varieties of English. Knowing that LLMs can be biased against certain varieties of English, as shown by Deas et al. (2023) and Srirag et al. (2025) for African-American English and Indian English respectively, little empirical evidence is found for their bias towards several other varieties. An exception is Multi-VALUE (Ziems et al., 2023) which creates synthetic datasets of language varieties using linguistically-informed syntactic transformations and reports a degraded performance on the GLUE benchmark. However, such synthetic texts do not accurately represent the spoken or written forms of the language variety because language varieties are more than syntax and encompass orthography, vocabulary, and cultural pragmatics (Nguyen, 2021). As a result, labelled datasets

Findings of the Association for Computational Linguistics: ACL 2025, pages 8413–8429
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

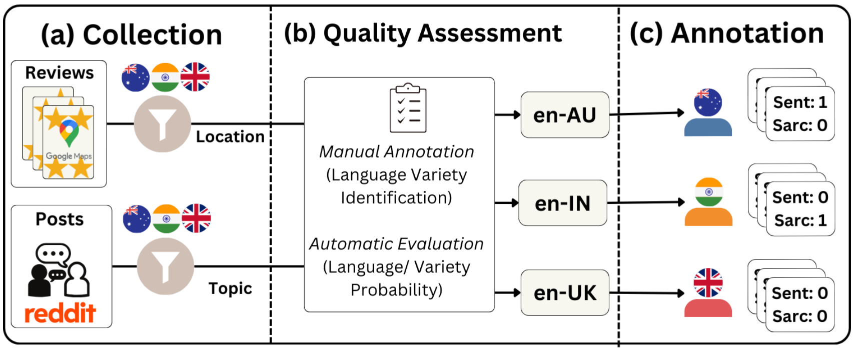


Figure 1: Creating dataset for BESSTIE; Collection in Section 2.1, Quality Assessment in Section 2.2, Annotation in 2.3.

of natural text are crucial to measure the bias in LLMs towards these varieties.

Acknowledging the role of the sentiment classification task (Socher et al., 2013) in GLUE, we introduce **BESSTIE**, A **BE**nchmark for **S**entiment and **S**arcasm classification for varie**TI**es of **E**nglish, namely, Australian (en-AU), Indian (en-IN), and British (en-UK). The relationship between sentiment and sarcasm classification as related tasks is well-understood in the context of sentiment analysis (Chauhan et al., 2020). **BESSTIE** comprises a manually labelled dataset of textual posts along with performance results of fine-tuned LLMs on the two classification tasks. The dataset consists of text samples collected from *two* web-based domains, namely, Google Places reviews, and Reddit comments, using *two* filtering methods: location-based and topic-based filtering respectively. We validate the representation of these varieties through a two-step process: (a) a manual annotation exercise and (b) automatic variety identification using models fine-tuned on the ICE-Corpora (Greenbaum and Nelson, 1996). Following that, we obtain manual annotations for boolean sentiment and sarcasm labels from native speakers of the language varieties.

We then use the dataset to model both the tasks as *binary classification* problems. We evaluate nine fine-tuned LLMs, spanning encoder/decoder and monolingual/multilingual models. Our results indicate that current models exhibit a significant decrease in performance when processing text written in en-IN, an outer-circle variety, in comparison with inner-circle varieties of English[1] (en-AU and en-UK). The performance on sarcasm classification is consistently low across the three varieties, providing pointers for future work. Following that, we examine the aggregate performance for the three varieties, the two tasks, and model attributes (encoder/decoder and monolingual/multilingual). We also examine cross-variety performance on all models highlighting the need for **BESSTIE**. **BESSTIE** evaluation provides a solid baseline to examine biases of contemporary LLMs, and will help to develop new methods to mitigate these biases. The novelty of the **BESSTIE** dataset can be seen in Table 1. **BESSTIE** improves not only upon popular sentiment/sarcasm-labelled datasets but also a recent dialectal benchmark, DialectBench (Faisal et al., 2024), which covers sentiment classification but for dialects of languages other than English. Therefore, the **BESSTIE** dataset is novel in the introduction of new language varieties (specifically, varieties of English) for both sentiment and sarcasm classification.

The contributions of this work are: (a) Creation of a manually annotated dataset of user-generated content with sentiment and sarcasm labels for three varieties of English, called **BESSTIE**; (b) Evaluation of nine LLMs on **BESSTIE** and analysing their performance to identify challenges arising due to language varieties.

2 Dataset Creation

Figure 1 shows the process followed to create a manually annotated dataset for sentiment and sarcasm classification.

2.1 Data Collection

We collect textual posts from two domains: Google Places reviews and Reddit comments, using two filtering methods: location-based and topic-based respectively. Location-based filtering implies that we select reviews that were posted for locations

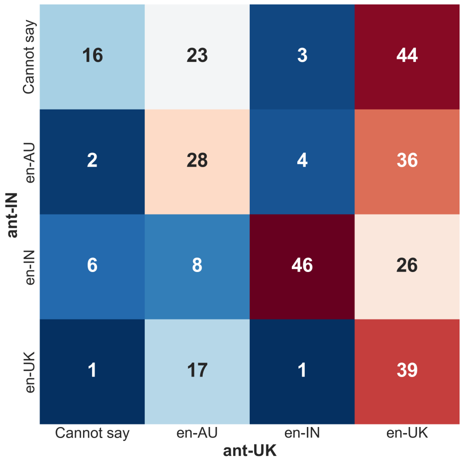
[1] Inner-circle English varieties are those spoken by people who use English as a first language. Outer-circle varieties are those spoken by bi/multilingual speakers.

in the three countries. Specifically, we collect reviews posted in cities in AU, IN and UK, and their corresponding ratings (1 to 5 *stars*, where 5 is the highest.) using the Google Places API[2] from the place types, also defined by the API. The criteria for city selection are based on population thresholds specific to each country (en-AU: 20K; en-IN:100K; en-UK: 50K). We filter out any non-English content using language probabilities calculated by fastText (Grave et al., 2018) word vectors, using a threshold of 0.98. Additionally, to mitigate the risk of including reviews written by tourists rather than residents, we exclude reviews from locations designated as *'tourist attractions'* by the Google Places API.

Variety	NAIVE	OURS
en-AU	0.97	0.79
en-IN	0.94	0.76
en-UK	0.96	0.81
μ	0.96	0.79

Table 2: Performance of DISTIL on sentiment classification with different labelling schemes. NAIVE refers to labels based on 1 and 5 *stars*, while OURS corresponds to labels based on 2 and 4 *stars*. μ denotes the average F-SCORE across all varieties

We also conduct a preliminary investigation using DistilBERT-Base (DISTIL;Sanh et al., 2020) to examine the influence of label semantics, the *star* ratings, on the performance of sentiment classification. We experiment with DISTIL on the said task using the collected Google Places reviews under two distinct labelling schemes: (a) extreme ratings, i.e., 1 and 5 *stars* (NAIVE); and (b) moderate ratings, i.e., 2 and 4 *stars* (OURS). Table 2 presents the performance of DISTIL for each variety under both schemes, reported using the macro-averaged F-SCORE. We observe that DISTIL yields strong performance when trained on NAIVE labels, achieving an average μ of 0.96 across the three varieties. However, task performance drops considerably with the use of the more ambiguous OURS labels, highlighting that they introduce increased difficulty and nuanced content. Following these findings, we filter the dataset to retain only reviews with intermediate ratings (2 and 4 *stars*), to better capture this nuance and avoid model overfitting to well-distinguished and polarised examples.



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

Figure 2: Confusion matrix showing the overlap in variety annotations between annotators, ant-IN, and ant-UK. Rows represent the labels assigned by ant-IN, and columns represent the labels assigned by ant-UK. The principle diagonal elements indicate agreement between annotators, while off-diagonal elements highlight disagreements.

To create the REDDIT subset, we employ topic-based filtering and choose up to four subreddits per variety (en-AU: *'melbourne', 'AustralianPolitics','AskAnAustralian'*; en-IN: *'India', 'IndiaSpeaks', 'BollyBlindsNGossip'*; en-UK: *'England', 'Britain', 'UnitedKingdom', 'GreatBritishMemes'*), where the topics are determined by native speakers of these language varieties. We select these subreddits based on the understanding that they feature discussions specific to a variety, making it highly likely that the post authors use the corresponding language variety. For each variety, we scrape $12,000$ comments evenly across the selected subreddits, capping at 20 comments per post and focusing on recent posts. These are then randomly sampled and standardised to $3,000$ comments per variety before manual annotation. We discard information such as user identifiers and post identifiers to maintain the anonymity of the user.

2.2 Quality Assessment

While previous studies utilise location-based filtering (Blodgett et al., 2016; Sun et al., 2023), we examine if the collected text from the two methods is indeed from these language varieties. To address this, we conduct two evaluations: (a) Manual annotation for national variety identification, and (b) Automated validation of national varieties.

[2] https://developers.google.com/maps/documentation/places/web-service/overview; Accessed on 10 February 2025

Variety	Subset	$P(\bar{\text{eng}})$	$P(\bar{\text{v}})$	F-SCORE
en-AU	GOOGLE	0.99	0.99	0.99
	REDDIT	0.98	0.95	0.93
en-IN	GOOGLE	0.99	0.94	0.91
	REDDIT	0.87	0.78	0.69
en-UK	GOOGLE	0.99	0.99	0.99
	REDDIT	0.98	0.93	0.90

Table 3: Data quality for each variety and subset. $P(\bar{\text{eng}})$ is the average language probability (calculated using fastText (Grave et al., 2018) word vectors) of a review or comment being in English. $P(\bar{\text{v}})$ represents the average variety probability of a review or comment being in the corresponding variety. F-SCORE measures the performance of the variety predictor.

Manual annotation We request en-IN and en-UK annotators, asking them to manually identify the variety of a given text. Specifically, they label a random sample of 300 texts (150 from reviews and 150 from comments) as *en-AU*, *en-IN*, *en-UK*, or *Cannot say*. Annotator agreement is measured using Cohen's kappa (κ).

With respect to the true label (based on the location or topic), results show higher agreement with the en-IN annotator (κ = 0.41) compared to the en-UK annotator (κ = 0.34), indicating fair to moderate consistency. The inter-annotator agreement itself is 0.26. Figure 2 shows that annotators find it difficult to agree on an inner-circle variety, and have the highest agreement when identifying the en-IN variety, with 46 agreements, demonstrating that the subset reliably represents this variety. As a result, the two subsets (reviews and comments) *collectively* form a good representative sample for language varieties.

Automated validation We use two predictors, a language predictor, and a variety predictor, to perform an automatic evaluation of our dataset. For language predictor, we use fastText word vectors and extract the probability of the text being in English. We then fine-tune DISTIL on ICE-Corpora, to use as a variety predictor. We model the task as binary classification (i.e., inner-circle vs. outer-circle variety classification). We use the ICE-Australia (Smith and Peters, 2023) and ICE-India subsets of the corpus[3], focusing on the transcriptions from unprompted monologues and private dialogues (S1A and S2A headers). Similar to language predictor, we extract the probability of text being in the corresponding variety.

$P(\bar{\text{eng}})$ in Table 3 is the probability that the text is written in English. $P(\bar{\text{v}})$ is the average variety probability. High values indicate that the variety predictor is capable of discerning text written in inner-circle varieties from those written in the outer-circle variety (i.e., en-IN). The lower language and variety probabilities on the REDDIT subset from en-IN, with 0.87 and 0.78 respectively, can be attributed to the code-mixed text in the subset. These results along with our manual annotation exercise show that the collected data represents the three language varieties. The probabilities from the model are in line with the ground truth, as measured using the F-SCORE.

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

Table 4: Inter-annotator agreement for the validation annotation exercise. Agreement is measured between the original annotator and an independent annotator for each language variety. Here Sent. and Sarc. denote sentiment and sarcasm labels respectively.

2.3 Annotation

We hire one annotator[4] each for the three language varieties. The annotators assign the processed reviews and posts two labels: sentiment and sarcasm. The choice of labels given are *negative*, *positive*, and *discard*. We instruct the annotators to use the *discard* label for uninformative examples with no apparent polarity and in a few cases, for computer-generated messages. The texts with *discard* label are discarded.

We conduct an additional annotation exercise to evaluate the reliability of the sentiment and sarcasm annotations. For this, we randomly selected 50 instances from each variety, annotated by the original annotators, and hire a new independent annotator for each variety. The inter-annotator agreement between in the original and the independent annotators, measured by κ, are reported in the Table 4.

[3] Due to lack of access to ICE-GB, we do not include en-UK variety in training the predictor.

[4] Two of these are the authors of the paper who were also involved in the evaluation.

Variety	Subset	Train	Valid	Test	% Pos. Sent.	% Pos. Sarc.	Avg. no. of words
en-AU	GOOGLE	946	130	270	73%	7%	63.97
	REDDIT	1763	241	501	32%	42%	51.72
en-IN	GOOGLE	1648	225	469	75%	1%	44.34
	REDDIT	1686	230	479	25%	13%	26.92
en-UK	GOOGLE	1817	248	517	75%	0%	72.21
	REDDIT	1007	138	287	12%	22%	38.04

Table 5: Dataset statistics for each variety, subset, split, and label type. % Pos. Sent. and % Pos. Sarc. indicates the proportion of samples with positive sentiment and true sarcasm respectively.

The absolute agreement values for sarcasm, though lower than those for sentiment, are comparable to prior work. In particular, Joshi et al. (2016) report a κ of 0.44 on their sarcasm labels. These scores suggest that both sentiment and sarcasm labels in our dataset are annotated with high degree of reliability. We compensate the annotators at the casual employment rate, equivalent to 22 USD per hour, as prescribed by the host institution. We provide the annotation guideline in Appendix A.

2.4 Dataset Statistics

The resultant **BESSTIE** dataset contains annotations for two tasks: sentiment and sarcasm classification, for the two domains and three varieties. Table 5 provides a breakdown of the annotated dataset (consisting of GOOGLE and REDDIT), including the number of samples in each split (Train, Validation, and Test), along with average word length. The % Pos. Sent. and % Pos. Sarc. indicates the proportion of samples with positive sentiment and true sarcasm respectively. These values are the same for the train, validation, and test set since we perform stratified sampling. We present example comments along with their corresponding sentiment and sarcasm labels in Appendix B. We also confirm that the text collected for Reddit subset of the **BESSTIE** is generated in 2024. While this subset represents a static snapshot, its recency allows it to reflect contemporary linguistic usage effectively.

3 Experiment Details

We conduct our experiments on a total of nine LLMs. These include six encoder models, *three* pre-trained on English corpora: BERT-Large (BERT) (Devlin et al., 2019), RoBERTa-Large (ROBERTA) (Liu et al., 2019), ALBERT-XXL-v2 (ALBERT) (Lan et al., 2020); and *three* multilingual models: Multilingual-BERT-Base (MBERT), Multilingual-DistilBERT-Base (MDISTIL) (Sanh et al., 2020), XLM-RoBERTa-Large (XLM-R) (Goyal et al., 2021). We also evaluate three decoder models, *one* pre-trained on English corpora: Gemma2-27B-Instruct (GEMMA) (Gemma Team et al., 2024) and *two* multilingual models: Mistral-Small-Instruct-2409 (MISTRAL) (Jiang et al., 2023), Qwen2.5-72B-Instruct (QWEN) (Yang et al., 2024). The model-specific details including architecture, language coverage of pre-training corpus, and number of parameters are described in Appendix C.

While encoder models are fine-tuned with full precision, we fine-tune quantised decoder models using QLoRA (Dettmers et al., 2023) adapters, targeting all linear layers. All models (including the variety predictor) are fined-tuned for 30 epochs, with a batch size of 8 and Adam optimiser. We choose an optimal learning rate by performing a grid search over the following values: 1e-5, 2e-5, and 3e-5. All experiments are performed using *two* NVIDIA A100 80GB GPUs.

We fine-tune encoder models using cross-entropy loss, weighted on class distribution (*positive*: '1' and *negative*: '0' labels). In contrast, for decoder models, we perform zero-shot instruction fine-tuning (Ouyang et al., 2022), using maximum likelihood estimation – a standard learning objective for causal language modeling (Jain et al., 2023).

For decoder models, we use two task-specific prompts, one for each task. For sentiment classification, we the prompt the model with:

> *"Generate the sentiment of the given text. 1 for positive sentiment, and 0 for negative sentiment. Do not give an explanation."*

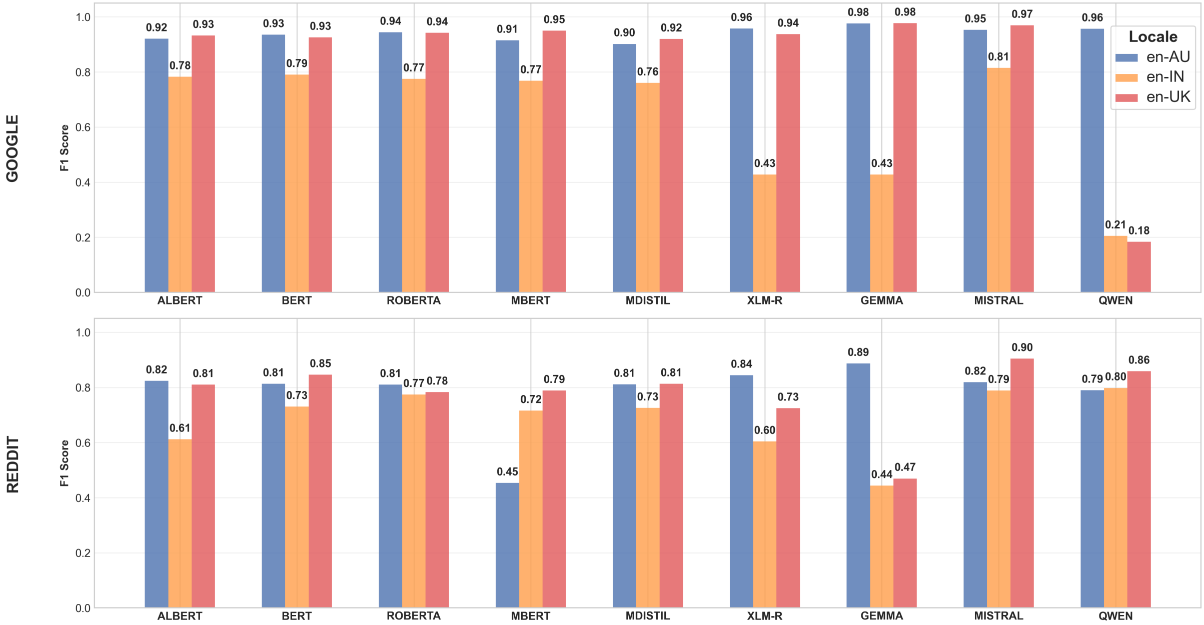


Figure 3: Performance comparison of various models on the sentiment classification task across different English varieties (en-AU, en-IN, and en-UK).

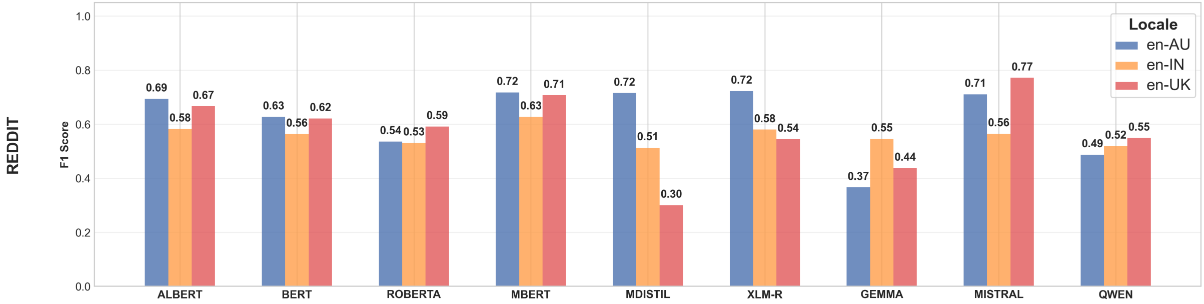


Figure 4: Performance comparison of various models on the sarcasm classification task across different English varieties (en-AU, en-IN, and en-UK).

Similarly, for sarcasm classification, we use:

> *"Predict if the given text is sarcastic. 1 if the text is sarcastic, and 0 if the text is not sarcastic. Do not give an explanation."*

Given the text and a task-specific prompt, the expected behaviour of the LLM is to generate either 1 (for *positive*) or 0 (for *negative*). The forms of these prompts are experimentally determined using a few test examples. Finally, we report our task performances on three macro-averaged F-SCORE, i.e., unweighted average, disregarding the imbalance in label distribution.

4 Results

We present our results and subsequent discussions to address the following questions: (a) How well do current LLMs perform on the benchmark tasks? (Section 4.1); (b) How do factors such as model properties and domain affect the task performance? (Section 4.2); (c) Can models trained on one language variety generalise well to others? (Section 4.3).

4.1 Task Performance

We first present our results on the two tasks where models are trained and evaluated on the same variety. Figures 3 and 4 describe the model performances, reported using F-SCORE, on the sentiment and sarcasm classification tasks respectively. While sentiment classification is performed using both GOOGLE and REDDIT subsets, due to the absence of sarcasm labels in GOOGLE subset (as shown in Table 5), sarcasm classification is conducted only on REDDIT subset. Table 6 shows that models, on average, perform better on both tasks when

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
μ	0.78	0.63	0.74

Table 6: Task performances, averaged across all models, on all varieties. μ represents the average performance across all domain-task pairs.

trained on en-AU variety, closely followed by performances on en-UK variety. Conversely, models consistently perform the worst on en-IN.

Sentiment Classification Table 7 shows that, for sentiment classification, MISTRAL achieves the highest average performance across all varieties, with an F-SCORE of 0.91 on the GOOGLE subset and 0.84 on the REDDIT subset. In contrast, QWEN reports the lowest average performance across all varieties on the GOOGLE subset, with an average F-SCORE of 0.45, while GEMMA exhibits the lowest average performance across all varieties on the REDDIT subset, with an F-SCORE of 0.60.

Sarcasm Classification Similarly, for sarcasm classification, both MISTRAL and MBERT achieve the highest average performance across all varieties, with an F-SCORE of 0.68, as shown in Table 7. In contrast, GEMMA again reports the lowest average performance across all varieties, with an average F-SCORE of 0.45. The lower performance on sarcasm classification as compared to sentiment classification is expected because sarcasm often relies on not just subtle linguistic cues but also localised contemporary contextual information (Abercrombie and Hovy, 2016).

4.2 Impact of Models and Domain

The average results for the models highlight the need to further probe into factors influencing model performances, namely, model properties and domain.

Model properties The results (μ values) in the corresponding columns of Table 7 describe the model performances (reporting F-SCORE), grouped based on model properties (encoder/decoder, mono/multilingual). The values show a consistent trend where encoder models report higher performance than decoder models across different tasks. This performance gap is expected as encoder architecture is inherently better suited for sequence classification tasks, while decoder architecture is optimised for text generation tasks. We also find that, although marginal, monolingual models report a higher average performance compared to multilingual models. This performance gap is prominent on the task involving GOOGLE subset, while multilingual models perform better on tasks involving REDDIT subset.

GOOGLE vs. REDDIT Our findings (the last row μ for the task columns in Table 7), indicate that models perform better on GOOGLE subset, achieving an F-SCORE of 0.81, compared to an F-SCORE of 0.75 on REDDIT subset. This difference can be explained by the distinct writing styles of the two sources. Reviews in GOOGLE are generally more formal and informative, while posts and comments from REDDIT often exhibit a short (as shown in Table 5), conversational tone typical of social media. Additionally, posts and comments in the REDDIT subset come from forums frequented by local speakers, which means the language and expressions used are more reflective of the variety and cultural nuances. We present results from cross-domain evaluation in the Appendix D.

4.3 Cross-variety Evaluation

Finally, we perform cross-variety evaluation. This refers to the scenario where the datasets used to train and test a model are from different groups (i.e., trained on en-AU, tested on en-IN, for example).

We report the cross-variety results of our best-performing model, MISTRAL. Figure 5 compares the performance of MISTRAL across three variants: pre-trained, in-variety fine-tuning, and cross-variety fine-tuning, for the three dataset subsets using colour-coded matrices. The rows indicate the subset that the model is trained on (i.e., en-AU, en-IN, and en-UK), along with PT, indicating that the pre-trained MISTRAL is used. Similarly, the columns indicate the test subset (i.e., en-AU, en-IN, and en-UK). The pre-trained variant of the model (the first row in each matrix) achieves high F-SCORE for en-AU and en-UK, for the sentiment classification task. Fine-tuning the model improves in-variety performance (shown by the principal diagonal of the matrices), with noticeable gains in F-SCORE on both subsets. The model performance on cross-variety evaluation remains relatively stable across different varieties of English, indicating that domain-specific data plays a minor role in in-

Model	Enco.	Mono.	GOOGLE-Sentiment	REDDIT-Sentiment	REDDIT-Sarcasm
ALBERT	✓	✓	0.88	0.75	0.65
BERT	✓	✓	0.88	0.80	0.60
ROBERTA	✓	✓	0.89	0.79	0.55
MBERT	✓	×	0.88	0.65	0.68
MDISTIL	✓	×	0.86	0.78	0.51
XLM-R	✓	×	0.77	0.73	0.62
GEMMA	×	✓	0.79	0.60	0.45
MISTRAL	×	×	0.91	0.84	0.68
QWEN	×	×	0.45	0.82	0.52
μ	✓: 0.74 ×: 0.67	✓: 0.72 ×: 0.71	0.81	0.75	0.59

Table 7: Model performances, averaged across all varieties, on sentiment and sarcasm classification tasks across two subsets (GOOGLE and REDDIT). μ represents the average performance across all models. 'Enco.' represents encoder models. 'Mono.' represents monolingual models.

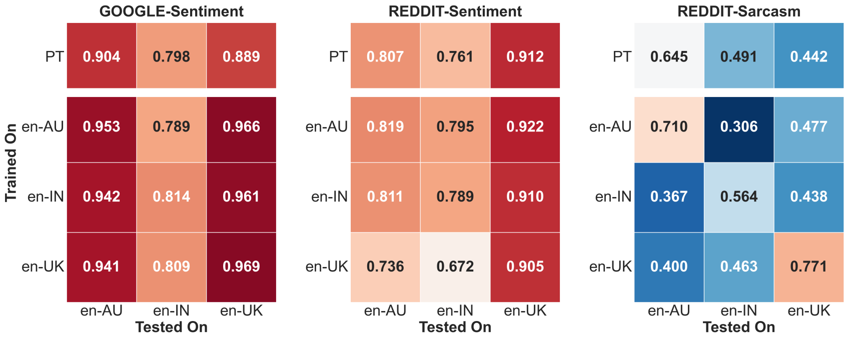


Figure 5: Cross-variety performance analysis of MISTRAL. The figure compares three different scenarios: pre-trained (PT), in-variety fine-tuning, and cross-variety fine-tuning for sentiment and sarcasm classification across all varieties.

fluencing cross-variety generalisation for sentiment classification.

The model performs significantly worse on sarcasm classification compared to sentiment classification. While in-variety evaluation shows that fine-tuning improves over the pre-trained model with substantial increases in F-SCORE, it negatively impacts cross-variety generalisation. This means there is a notable decrease in F-SCORE across other varieties when the model is fine-tuned on a specific variety. This supports the idea that sarcasm classification requires the model to grasp linguistic and cultural nuances specific to a variety, making generalisation across varieties challenging.

5 Error Analysis

We randomly sample up to 30 examples misclassified by MISTRAL, from each variety and domain-task configuration. We then manually analyse these misclassified examples to identify pervasive or obligatory dialect features, as defined by eWAVE (Kortmann et al., 2020). A dialect feature is defined as any lexical or syntactic variation from standard English. In addition, we identify other features, including locale-specific colloquial expressions, instances that require additional context, and occurrences of code-mixed text. Table 8 summarises the sample sizes and the counts of these identified features for each variety. Detailed ex-

Variety	Sample	DIAL	COLL	CONT	CODE
en-AU	70	9	28	6	-
en-IN	90	97	33	3	8
en-UK	53	7	15	5	-

Table 8: Counts of identified features in misclassified examples from MISTRAL for each variety. DIAL denotes dialect features, COLL denotes locale-specific colloquial expressions, CONT denotes instances requiring additional context, and CODE denotes occurrences of code-mixed text.

amples for each feature type are provided in Appendix E.

The en-IN variety exhibits a high count of dialect features (97) relative to its sample size (90), suggesting that regional variations significantly challenge model performance. Locale-specific colloquial expressions are prevalent in all varieties (en-AU: 28; en-IN: 33; en-UK: 15), highlighting the impact of regional lexicons on model performance. These findings motivate future research in sentiment and sarcasm classification to better accommodate such colloquial expressions for inner-circle varieties, and to additionally adapt to dialect features for outer-circle varieties of English.

6 Related work

NLP Benchmarks such as GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), and DynaBench (Kiela et al., 2021) have played a pivotal role in evaluating LLMs on multiple NLP tasks, including sentiment classification. However, these benchmarks are limited in their ability to capture the linguistic nuances of non-standard language varieties. Efforts to address this include datasets for African languages and dialects (Muhammad et al., 2023), Arabic dialects (Elmadany et al., 2023), and CreoleVal (Lent et al., 2024) which provides datasets for 28 creole languages. More recent progress includes DialectBench (Faisal et al., 2024), which presents a collection of dialectal datasets and associated NLP models across 281 dialectal varieties for several tasks. However, the benchmark does not contain English dialectal datasets for sentiment and sarcasm classification.

Similarly, several sarcasm classification datasets for standard variety of English exist, consisting of reddit comments (Khodak et al., 2018), amazon reviews (Filatova, 2012), and tweets (Ptáček et al., 2014; Painter et al., 2022; Abercrombie and Hovy, 2016; Oprea and Magdy, 2020). While datasets like ArSarcasm-v2 (Abu Farha et al., 2021) present annotations for sarcasm, sentiment, and the dialect of each tweet for Arabic, there is a notable absence of sarcasm classification datasets that account for non-standard varieties of English. Similar to their work, we also create a dataset that contains both sentiment and sarcasm labels. To the best of our knowledge, **BESSTIE** is the first benchmark for sentiment and sarcasm classification specifically for varieties of English.

7 Conclusion

This paper presents **BESSTIE**, a benchmark for sentiment and sarcasm classification across three varieties of English: Australian (en-AU), Indian (en-IN), and British (en-UK). Our evaluation spans nine LLMs, six encoders and three decoders, assessed on datasets collected from Google Places reviews (GOOGLE) and Reddit comments (REDDIT). The models perform consistently better on en-AU and en-UK (i.e., inner-circle varieties) than on en-IN (i.e., the outer-circle variety) for both tasks. We verify the variety representation of collected data using: (a) manual annotation and (b) automated validation. We also perform an additional annotation exercise to verify the reliability of the annotations. Although models report high performance on sentiment classification (F-SCORE of 0.81 and 0.75, for GOOGLE and REDDIT respectively), they struggle with detecting sarcasm (F-SCORE of 0.59 on REDDIT), indicating that sarcasm classification is still largely unsolved. This is supported by our cross-variety evaluation which reveals limited generalisation capability for sarcasm classification, where cultural and contextual understanding is crucial. Notably, monolingual models marginally outperform multilingual models, suggesting that language diversity in pre-training does not extend to varieties of a language. Our error analysis provides potential considerations to drive future research in sentiment and sarcasm classification to better accommodate colloquial expressions for inner-circle varieties, and dialect features for outer-circle varieties of English. In conclusion, **BESSTIE** provides a resource to measure bias in LLMs towards non-standard English varieties for the two tasks.

Limitations

Our assumption of national varieties as representative forms of dialect is a simplification. Within each locale, there are significant regional, sociolectal, and generational differences. For example, Australian English spoken in Sydney may differ significantly from that in Perth. The language used in online communications evolves rapidly, driven by cultural trends, viral content, and changing societal norms. Phrases, slang, and expressions can gain popularity and fade quickly. A static dataset, such as ours, is hence limited in reflecting these continuous and dynamic changes in language use. While our efforts to remove non-English text were thorough, there may still be instances where language mixing or code-switching occurs, especially in the Indian English subset, affecting model performance. The annotation process was limited by the number of human annotators and their subjective interpretations. While we perform an additional annotation exercise to validate the annotation process, use of a single annotator can introduce individual biases. We also note that comments from Reddit may differ from other social media sites, such as Twitter/X, in terms of user demographics and discourse style. This potentially influences the nature of topic-based discussions captured in our dataset. Our reliance on Reddit data is due to cost-related constraints associated with accessing Twitter/X API.

Ethical Considerations

The project received ethics approval from the Human Research Ethics Committee at UNSW, Sydney (reference number: iRECS6514), the host organisation for this research.We treat the reviews as separate texts and do not aggregate them by user in any way. The Reddit comments and Google reviews used in this paper are from publicly available posts, accessible through official APIs. We adhere to the terms of service and policies outlined by these platforms while gathering data. We do not attempt to de-identify any users or collect aggregate information about individual users. The dataset, including sentiment and sarcasm annotations, is shared in a manner that preserves user privacy and abides by the rules and guidelines of the source platforms.

Acknowledgements

This work was funded by Google's Research Scholar grant, awarded to Aditya Joshi and Diptesh Kanojia in 2024.

References

Gavin Abercrombie and Dirk Hovy. 2016. Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of Twitter conversations. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 107–113, Berlin, Germany. Association for Computational Linguistics.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Basma Alharbi, Hind Alamro, Manal Alshehri, Zuhair Khayyat, Manal Kalkatawi, Inji Ibrahim Jaber, and Xiangliang Zhang. 2020. Asad: A twitter-based benchmark arabic sentiment analysis dataset. *arXiv preprint arXiv:2011.00578*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4351–4360.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain. Association for Computational Linguistics.

Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. ORCA: A challenging benchmark for Arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 392–398, Istanbul, Turkey. European Language Resources Association (ELRA).

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.

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)*.

Sidney Greenbaum and Gerald Nelson. 1996. The international corpus of english (ice) project. *World Englishes*, 15(1):3–15.

Nihal Jain, Dejiao Zhang, Wasi Uddin Ahmad, Zijian Wang, Feng Nan, Xiaopeng Li, Ming Tan, Ramesh Nallapati, Baishakhi Ray, Parminder Bhatia, Xiaofei Ma, and Bing Xiang. 2023. ContraCLM: Contrastive learning for causal language model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6436–6459, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.*, 57(6).

Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, and Mark J. Carman. 2016. Harnessing sequence labeling for sarcasm detection in dialogue from TV series 'Friends'. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 146–155, Berlin, Germany. Association for Computational Linguistics.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Bernd Kortmann, Kerstin Lunkenheimer, and Katharina Ehret, editors. 2020. *eWAVE*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, Anders Søgaard, and Johannes Bjerva. 2024. Creoleval: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages.

Dong Nguyen. 2021. 10 dialect variation on social media. *Similar Languages, Varieties, and Dialects: A Computational Perspective*, page 204.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jordan Painter, Helen Treharne, and Diptesh Kanojia. 2022. Utilizing weak supervision to create S3D: A sarcasm annotated dataset. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 197–206, Abu Dhabi, UAE. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on Czech and English Twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 213–223, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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.

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

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.

Dipankar Srirag, Nihar Ranjan Sahoo, and Aditya Joshi. 2025. Evaluating dialect robustness of language models via conversation understanding. In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 24–38, Abu Dhabi. Association for Computational Linguistics.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. Dialect-robust evaluation of generated text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6028, Toronto, Canada. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

A Annotation Guidelines

Task Overview

Your task involves determining whether each Google review or Reddit comment expresses positive or negative sentiment and whether it uses sarcasm.

Annotation Instructions

For Sentiment Classification:

- Positive Sentiment (Label 1): Annotate a comment as positive if it reflects favorable emotions such as happiness, satisfaction, agreement, or excitement.
- Negative Sentiment (Label 0): Annotate a comment as negative if it conveys unfavorable emotions such as sadness, disappointment, frustration, or criticism.
- Neither Positive nor Negative (Discarded, Label 2): discard the comment if it does not convey clear positive or negative sentiment (mark the same sarcasm label as 2).

For Sarcasm Classification:

- Sarcastic Comment (Label 1): Annotate if the comment uses irony or mockery to express contempt or ridicule.
- Non-Sarcastic Comment (Label 0): Annotate if the comment is straightforward and does not employ sarcasm or irony.

Additional Guidelines

Consider the tone and context of the comment when assigning sentiment or sarcasm labels. For sarcasm, look for indicators such as exaggerated language, contradictions, or unexpected statements. Base your decision on the sentiment explicitly expressed in the text, avoiding personal biases. It is acceptable to mark a comment as undecided or seek clarification if you cannot confidently assign a sentiment label.

B Dataset Examples

Variety	Comment	Sentiment	Sarcasm
en-AU	Well set out, good stock and very friendly country staff	1	0
en-AU	S**t. Christ isn't back, is he?	0	1
en-IN	Good quality foods are supplied here	1	0
en-IN	Coz we all have free internet.	0	1
en-UK	Traditional friendly pub. Excellent beer	1	0
en-UK	What a brave potatriot	0	1

Table 9: Reviews and comments from different varieties with their corresponding annotated sentiment and sarcasm labels.

The examples presented in Table 9 illustrate the diversity of language use and the challenges involved in detecting both sentiment and sarcasm across the varieties.

C Additional Model Details

We describe additional model details, including architecture, the language coverage in their pre-training data, and the number of trainable parameters in Table 10.

Model	Arch.	Lang.	# Params
ALBERT	Encoder	English	223 M
BERT	Encoder	English	340 M
ROBERTA	Encoder	English	355 M
MBERT	Encoder	Multilingual	177 M
MDISTIL	Encoder	Multilingual	134 M
XLM-R	Encoder	Multilingual	355 M
GEMMA	Decoder	English	27 B
MISTRAL	Decoder	Multilingual	22 B
QWEN	Decoder	Multilingual	72 B

Table 10: Details of models including architecture (Arch.), language coverage in pre-training data (Lang.), and number of trainable parameters (# Params). Here, M is millions and B is billion.

D Cross-domain Evaluation

We conduct experiments with domain data from GOOGLE, and REDDIT, to assess the cross-domain robustness of our sentiment classification models. Models trained on one domain are tested on the other domain, and vice-versa, for each of the three varieties of English. Figure 7 shows that models perform better when trained on in-domain data, for both MISTRAL and BERT. We observe that pre-trained MISTRAL performs better for all varieties in comparison to results from the cross-domain experiments. The corresponding results for BERT are in Figure 6.

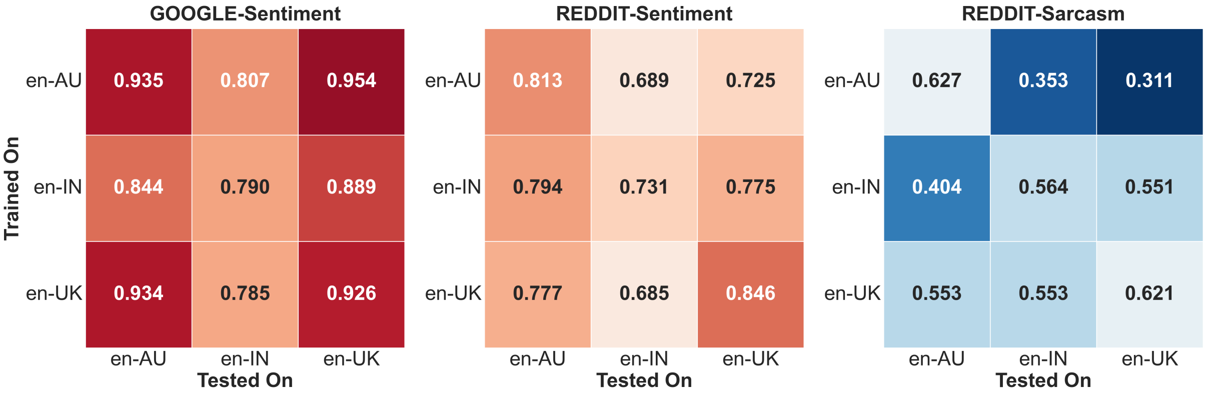


Figure 6: Cross-variety performance analysis of BERT. The figure compares two different scenarios: in-variety fine-tuning, and cross-variety fine-tuning for sentiment and sarcasm classification across all varieties.

While MISTRAL outperforms BERT at both in- and cross-domain experiments, there is significant variation in cross-domain results, highlighting the challenge in *generalization*, given domain variance in fine-tuned data. Moreover, the direction of cross-domain experiments also affects performance. Both models perform better for REDDIT → GOOGLE, suggesting that *models may be better at transferring sentiment from a relatively informal domain*. The overall difference in cross-domain performance is significantly higher for inner-circle varieties, compared to en-IN, indicating that fine-tuning with mixed-domain data may be less effective for such varieties. Across the varieties, however, en-UK, an inner-circle variety, performs better for GOOGLE → REDDIT, whereas for the BERT model, this is observed with the en-IN variety.

E Detailed Feature Types

We present detailed feature types identified in misclassified samples from different English varieties using the MISTRAL model. Each table lists the feature, an example illustrating the feature, and the frequency

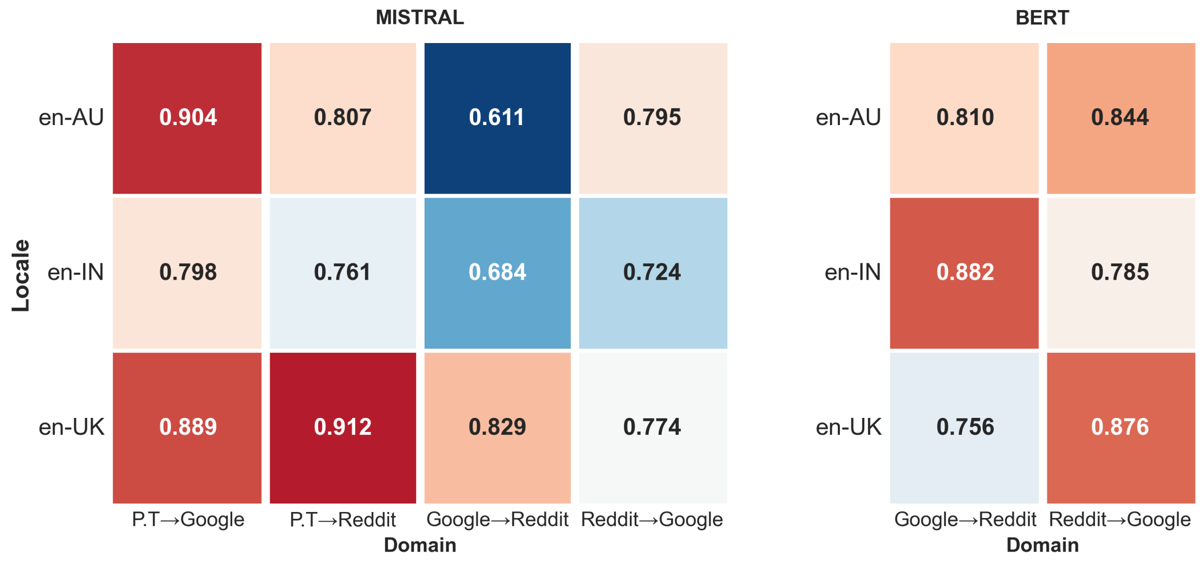


Figure 7: Cross-domain performance analysis of MISTRAL and BERT comparing a decoder and an encoder model.

count. Table 11 shows features found in en-AU samples, Table 12 shows features found in en-IN samples, and Table 13 shows features found in en-UK samples.

Feature	Example	Count
DIAL		
LEVELLING OF PAST TENSE VERB FORMS	*when we order passionfruit cheecake and never get it*	2
PRONOUN DROP	*(We) also ordered 2 noodles, fish and dessert*	3
THERE'S WITH PLURAL SUBJECTS	*There's been plenty of fossils and opals uncovered*	1
NO AUXILIARY IN YES/NO QUESTIONS	*(Do) people still watch commercial TV?*	2
COLL	*Hi Hospo workers afternoon when you wake up*	28
CONT	*Beige night*	6

Table 11: Features identified in the en-AU samples misclassified by MISTRAL.

Feature	Example	Count
DIAL		
ARTICLE OMISSION	*Kamal is playing (a) cameo role*	35
PRONOUN DROP	*(I) can't wait for animal park man*	22
OBJECT/SUBJECT FRONTING	*Laws of India doesn't apply on white man, he is above everything*	12
'VERY' AS QUALIFIER	*quantity is very very less*	10
COPULA OMISSION	*(I am) Waiting for cash to arrive in a tempo*	18
COLL	*It means benstokes*	33
CONT	*Alia Advani as her earlier name suits better.*	3
CODE	*Bridge chori ho jaaega*	8

Table 12: Features identified in the en-IN samples misclassified by MISTRAL.

Feature	Example	Count
DIAL		
WAS FOR CONDITIONAL WERE	*if it was in the autumn.*	1
GROUP GENITIVE	*their and their old boy network*	1
LEVELLING OF PAST TENSE VERB FORMS	*but could been more spicy.*	1
PRONOUN DROP	*(It) took just 2 and a half minutes to lose his head*	4
COLL	*Warra calm and coherent interview.*	15
CONT	*Haha yup!*	5

Table 13: Features identified in the en-UK samples misclassified by MISTRAL.