PennState College of Information Sciences and Technology

DIA-HARM: Dialectal Disparities in Harmful Content Detection Across 50 English Dialects

Jason Lucas[1], Matt Murtagh[*2], Ali Al-Lawati[*1], Uchendu Uchendu[1], Adaku Uchendu[3], Dongwon Lee[1]

[1] Pennsylvania State University, PIKE Research Lab · [2] Trinity College Dublin · [3] MIT Lincoln Laboratory · **Equal contribution*

195K+	50
Dialectal Samples	English Dialects
16	**9**
Detection Models	Source Benchmarks

OVERVIEW

Harmful content detectors are predominantly developed and evaluated on **Standard American English (SAE)**, leaving their robustness to dialectal variation unexplored. We present **DIA-HARM**, the first benchmark for evaluating disinformation detection robustness across **50 English dialects** spanning U.S., British, African, Caribbean, and Asia-Pacific varieties. Using Multi-VALUE's linguistically-grounded transformations, we introduce **D³** (Dialectal Disinformation Detection), a corpus of **195K+ samples** from 9 established benchmarks. Our evaluation of 16 detection models reveals systematic vulnerabilities: human-written dialectal content degrades detection by **1.4–3.6% F1**, while AI-generated content remains stable. Models with multilingual pre-training substantially outperform monolingual encoders and zero-shot LLMs fail catastrophically. These findings demonstrate that current detectors may **systematically disadvantage hundreds of millions of non-SAE speakers**.

EVALUATION PIPELINE

SAE Dataset
1. F3
2. LIAR
3. CoAID
4. MM-COVID
5. FakeNewsNet
6. MeDi
7. MultiClaim
8. Twitter15
9. Twitter16

→

Multi-VALUE
189 Morphosyntactic Rules
- · Pronouns
- · Noun Phrase
- · Tense & Aspect
- · Modal Verbs
- · Verb Morph.
- · Negation
- · Agreement
- · Relativization
- · Complement.
- · Adverbials
- · Word Order
- · Discourse

12 Grammatical Categories · eWAVE

→

D-PURIFY
Quality Validation

BERTScore	>0.50
BARTScore	≥−7.5
AlignScore	logical
METEOR	≥0.40
ROUGE-L	(0.01,1.0)
DiffLib	[0.01,0.99]
Feat. Accuracy	eWAVE

93.7% retention · 194,960 samples

→

D³ Corpus
195K+ Dialectal Samples
- 🇺🇸 U.S. Varieties (9)
- 🇬🇧 British/UK (11)
- 🌍 Africa (11)
- 🌏 Asia-Pacific (12)
- 🌊 Caribbean/Atlantic (5)
- 🌐 Other (2)

50 Dialects Total

→

Evaluation
16 Detection Models

Fine-Tuned (10)
mDeBERTa · mBERT · BERT · RoBERTa · DeBERTa · XLM-R · CT-BERT · BiGRU · TextCNN · dEFEND

Zero-Shot (6)
Mistral-7B · Llama-3.1-8B · Llama-3.2-1B · Gemma-3-1B · Qwen3-8B · Qwen3-4B

SQ1 Unseen · SQ2 Seen · SQ3 Transfer · SQ4 Arch.

DIALECT COVERAGE — 50 ENGLISH VARIETIES ACROSS 5 GEOGRAPHIC REGIONS

U.S. (9)
AAVE · Appalachian · Chicano · Ozark · Newfoundland · Colloquial · SE Enclave

British/UK (11)
Scottish · Irish · Welsh · Manx · Channel Islands · N/SE/SW/EA · Orkney & Shetland

Africa (11)
Nigerian · Ghanaian · Cameroon · Kenyan · Ugandan · Tanzanian · S. African (3) · Liberian

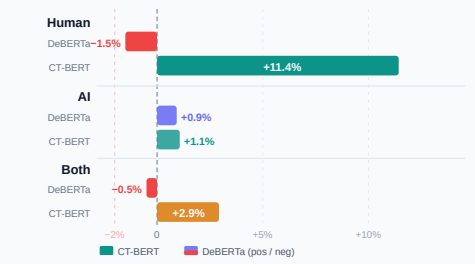
Asia-Pacific (12)
Indian · Pakistani · Sri Lankan · Singlish · Malaysian · Philippine · HK · Australian (2) · NZ · Fiji (2)

Caribbean/Atlantic (5)
Jamaican · Bahamian · Tristan da Cunha · Falkland Islands · St. Helena

KEY FINDINGS

SQ1 SAE → Unseen Dialects

ΔF1 VS SAE BASELINE · CT-BERT VS DEBERTA



−1.5%	+11.4%	+2.9%
DeBERTa Human	CT-BERT Human	CT-BERT Both

SQ2 Dialect-Aware Training

Model	SAE-only	Dia-only	SAE+Dia
RoBERTa	87.3%§	97.1%▲	96.9%
mDeBERTa†	97.0%	97.0%	96.8%
BERT-Large	97.2%	96.9%	95.8%▼
dEFEND	94.8%	88.6%▼	87.5%▼
XLM-R†	83.3%§	85.4%§	79.7%§

† Multilingual · §Failure (<90%) · ▲Best · ▼Worst · **SAE anchoring paradox:** including SAE data harms traditional models

SQ3 Cross-Dialectal Transfer

2,450 dialect pairs evaluated

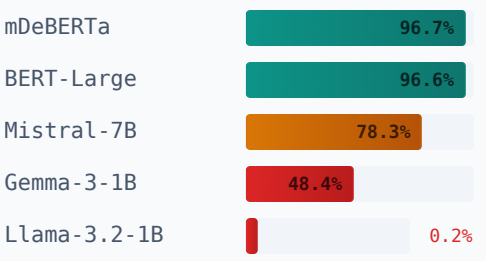
mDeBERTa **97.2% avg** vs **39.8%**

mDeBERTa ↔ XLM-R

Best sources: Ghanaian (92.5%), Manx (91.6%), Tristan da Cunha (91.0%)
Hardest targets: Maltese (80.0%), Australian Vernacular (80.1%), SE England (80.1%)

Transfer is **asymmetric**: Scottish & Welsh are poor sources but easy targets.

SQ4 Fine-Tuned vs Zero-Shot



18–96 F1-pt gap between fine-tuned and zero-shot. Up to **98% abstention** in smaller LLMs on dialectal inputs.

MODEL ARCHITECTURE PERFORMANCE

	MODEL	HUMAN F1	AI F1	ΔSAE
E	mDeBERTa†	96.7%	99.4%	−1.5
E	CT-BERT	97.2%	99.4%	+11.4
E	BERT-Large	96.6%	99.2%	−1.5
E	mBERT†	96.6%	99.4%	−1.4
E	RoBERTa	96.3%	99.5%	−1.5
E	DeBERTa	94.7%	97.4%	−1.5
E	dEFEND	93.9%	98.3%	−3.6
D	Mistral-7B	78.3%	78.3%	−11.0
D	Llama-3.1-8B	67.2%	67.2%	−20.1
D	Gemma-3-1B	48.4%	48.4%	−27.4
D	Qwen3-8B	26.6%	26.6%	−11.4
D	Llama-3.2-1B	0.2%	0.2%	−1.9

E Encoder (Fine-Tuned) D Decoder (Zero-Shot)

† Multilingual pre-training. Human F1 = dialect avg on human content.

ASYMMETRIC HARM ANALYSIS

⚠️ **Over-Flagging (False Positives)** +8.3%
Authentic dialectal speech silenced as disinformation. Dominates 6.5:1 over under-protection (27,020 FPs vs 4,169 FNs).

🛡️ **Under-Protection (False Negatives)** +1.4%
Disinformation evades detection in dialectal form. 33/50 dialects under-protected under unseen conditions.

🔄 **Training Composition Reversal** Δ flip
RoBERTa flips from over-flagging (dialect-only) to under-protection (SAE-anchored). *Who is harmed depends on training choices.*

81.4% of FPs & 75.5% of FNs made at >95% confidence
RoBERTa: 99.5% mean FP confidence — rules out calibration fixes. Dialectal features are *encoded as class-discriminative signals*.

RECOMMENDATIONS FOR PRACTITIONERS

R1 **Prefer multilingual encoders.** mDeBERTa and mBERT consistently achieve superior dialectal robustness — 97.2% avg F1 across 2,450 transfer pairs. Cross-lingual pre-training induces dialect-robust representations.

R2 **Adopt dialect-diverse fine-tuning without SAE anchoring.** Dialect-only training recovers failures (RoBERTa: 87.3%→97.1%) and avoids feature collapse toward standard patterns.

R3 **Conduct pre-deployment dialectal auditing.** Use D³ benchmark across dialect families with minimum thresholds per region — not just aggregate metrics.

R4 **Avoid zero-shot LLMs for content moderation.** 18–97 F1-pt gaps and abstention rates up to 98% render zero-shot approaches unsuitable for diverse communities.

R5 **Monitor dialectal performance longitudinally.** Retraining on SAE-dominant data may reintroduce bias — integrate continuous dialectal benchmarking into update pipelines.

DIA-HARM
D³ corpus · Evaluation scripts · Model checkpoints

🌐 jasonslucas.com | GitHub: github.com/jsl5710/dia-harm | jsl5710@psu.edu

Pennsylvania State University · PIKE Research Lab
Trinity College Dublin · MIT Lincoln Laboratory