SLANG-GraphRAG: Multi-Layered Retrieval with Domain-Specific Knowledge for Low Resource Social Media Conversations

Ifeoluwa Wuraola[1], Daniel Marciniak[1], Nina Dethlefs[2]
[1]University of Hull, UK
[2]Loughborough University, UK
i.a.wuraola-2021@hull.ac.uk, d.f.marciniak@hull.ac.uk, n.dethlefs@lboro.ac.uk

Abstract

Emotion classification on social media is especially difficult when texts include informal, culturally grounded language like slang. Standard NLP benchmarks often miss these nuances, particularly in low-resource settings. We present SLANG-GraphRAG, a retrieval-augmented framework that integrates a culture-specific slang knowledge graph into large language models via one-shot prompting. Using multiple retrieval strategies, we incorporate slang definitions, regional usage, and conversational context. Our results show that incorporating structured cultural knowledge into the retrieval process leads to significant improvements, improving accuracy by up to 31% and F1 score by 28%, outperforming traditional and unstructured retrieval methods. To better evaluate model behavior, we propose a probabilistic metric that reflects the distribution of human annotations, providing a more nuanced measure of performance. This highlights the value of culturally sensitive applications and more balanced evaluation in subjective NLP tasks.

1 Introduction

Recent advancements in natural language processing (NLP) have led to high-performing models across various tasks (Yao et al., 2023). However, a significant gap remains between the capabilities of these models and the linguistic realities of global users, particularly those from low-resource language communities (AlKhamissi et al., 2024; Shu et al., 2024; Ipa et al., 2025). Most state-of-the-art models are trained on large, Anglocentric datasets (Conneau et al., 2018), limiting their ability to handle the informal, multilingual, and culturally embedded language typical of social media. This includes specific dialects or slang and code-switching that are essential to conveying emotion, sentiment, and meaning in these contexts. This limitation is pronounced especially in emotion classification tasks, where subtle, culture-specific cues are important. For instance, models scoring high on benchmarks may miss the emotional tone in climate protest tweets written in Nigerian Pidgin (Wuraola et al., 2024). Similarly, Oyewusi et al. (2020) show that standard sentiment models misinterpret Nigerian Pidgin expressions such as "ginger" for motivation or "tank" for gratitude. Althobaiti (2023) further demonstrates that even strong baseline models under-perform on Arabic emotion classification tasks. This limitation emphasizes the need for architectures tailored to these language challenges.

A notable development is Retrieval Augmented Generation (RAG) (Lewis et al., 2020) which enhances pre-trained language models by integrating external information and retrieval components into task workflows (Leng et al., 2024; Gao et al., 2024). The RAG framework improves the efficiency of traditional large language models (LLMs), addressing issues like hallucinations (Chen et al., 2024; Lewis et al., 2020), inaccurate responses, and the need for faster information delivery (Ahmad, 2024). Previous studies on traditional text-based RAGs have demonstrated positive results and performances. However, they often struggle in domains with unstructured data, failing to capture the relational and contextual understanding needed for queries that require reasoning across an entire dataset, such as identifying trends or patterns over time (Edge et al., 2025). To address these challenges, graph-based RAG frameworks use knowledge graphs (KG) to encode entities and relationships in a structured, interpretable form (Wan et al., 2025), making them well-suited for capturing the complex, relational nature of social media data (Hu et al., 2024).

In this paper, we present SLANG-GraphRAG, a Knowledge Graph-based RAG framework designed to handle informal, culturally grounded, low-resource social media text, with a focus on culturally specific expressions from Nigeria and the UK. Our approach combines slang dictionaries, region-specific definitions, and sociocultural

Findings of the Association for Computational Linguistics: EACL 2026, pages 3919–3931

context through knowledge graphs, allowing meaningful interpretation of non-standard expressions. We integrate this framework into emotion classification tasks by fine-tuning and evaluating LLMs on culturally nuanced and emotion-rich datasets, demonstrating improvements over existing state-of-the-art approaches. In summary, our key contributions include:

- We develop a multilayered graph-based RAG architecture that integrates informal linguistic signals, including slang, into the generation process.
- We construct a domain-specific knowledge graph linking slang terms to paraphrases and contextual information, enhancing interpretability and grounding.
- We integrate conversational patterns into the classification process using a knowledge graph, boosting relevance in informal and culturally nuanced conversations.
- We propose a novel evaluation framework for informal text rich in emotions using normalized quantitative metrics benchmarked against human annotations, emphasizing cross-cultural sensitivity.

By focusing on informal and emotionally expressive language, particularly in non-Anglocentric regions, SLANG-GraphRAG supports a movement toward NLP models that are technically robust, inclusive, interpretable, and responsive to diverse online communication.

2 Related Works

NLP for Slang and Informal Language Recent NLP research has increasingly addressed slang and informal language. Wilson et al. (2020) introduced embeddings trained on Urban Dictionary for sentiment and sarcasm detection. Ni and Wang (2017) proposed a dual-encoder model generating contextual explanations for non-standard expressions. Kamath et al. (2024) showed that FastText embeddings incorporating Gen Z slang improve sentiment analysis accuracy. Sun et al. (2024) used LLMs and fine-tuned BERT models to detect slang and attribute regional and historical context, highlighting the role of cultural knowledge. Keidar et al. (2022) analyzed slang evolution, finding less semantic change but greater frequency shifts compared to standard words.

RAG applications in NLP We review RAG applications in NLP across two areas: (1) Text-based RAG, which retrieves from unstructured documents but may introduce redundancy and hallucinations in handling implicit knowledge; and (2) Graph-based RAG, which leverages structured entity-relation graphs for better reasoning and context. Both approaches have enhanced LLM performance in tasks like question answering (Lewis et al., 2020; Alshammary et al., 2024), dialogue generation (Cai et al., 2024), and summarization (Han et al., 2025). For example, Lu and Cosgun (2025) enriched domain content using Wikipedia, while Liu et al. (2024) used RAG to support context-aware learning in AI-assisted learning.

Recent studies have compared RAG and GraphRAG frameworks, with Han et al. (2025) systematically evaluating their performance on benchmark tasks. For instance, Edge et al. (2025) introduce GraphRAG to improve holistic document understanding, while Dong et al. (2024) propose G-RAG, which incorporates graph reranking to refine retrieval relevance. SubgraphRAG (Li et al., 2025) optimizes knowledge selection for smaller models, and Graphusion (Wang et al., 2025) enhances entity and relation modeling via graph-guided extraction. HiTA (Liu et al., 2024) applies RAG to improve educational assistance through better knowledge contextualization.

Cross-cultural performance of RAG frameworks While most RAG systems in high-resource settings have demonstrated great performance, their application in low-resource settings presents unique challenges and opportunities (Shandilya and Palmer, 2025). These challenges include sparse linguistic resources, limited training data, and greater linguistic diversity, often with informal or dialectal variations that are difficult to model (Zhong et al., 2024). Standard retrievers often struggle with the lexical and morphological variations of underrepresented languages, impacting retrieval accuracy and the quality of generated content.

To address this, recent studies have adapted RAG for multilingual and low-resource scenarios using techniques such as cross-lingual transfer, multilingual pretraining, and culturally grounded retrieval. For instance, Li et al. (2023) applied cross-lingual retrieval to Bangla, while Wang et al. (2024) introduced ReMaKE for multilingual knowledge editing. Tailored solutions have also emerged for Ara-

bic (Alshammary et al., 2024), Swahili (Ndimbo et al., 2025), and Albanian (Ramadani and Doko, 2025), emphasizing the importance of culturally and linguistically adaptive RAG frameworks. Also, in emotion-based tasks, recent studies have integrated RAG to enhance emotional understanding. Vologina et al. (2024) improved emotional expressiveness in Russian dialogue agents while Ngo-Ho et al. (2025) applied RAG to Vietnamese speech comprehension, highlighting the role of cultural context in emotion recognition within low-resource settings.

While both text-based and graph-based RAG frameworks have advanced significantly, the integration of knowledge through structured graphs offers a powerful approach for cross-cultural NLP. Prior work has largely centered on formal, Anglocentric ontologies like Wikidata, which fail to capture informal, culturally embedded language in low-resource settings. Despite the potential of graph-based RAG, its application to non-Anglocentric contexts remains underexplored. This work addresses that gap by applying graph-RAG to slang-rich, emotion-laden social media conversations.

Our model, SLANG-GraphRAG, constructs a multi-layer knowledge graph that links slang terms to their meanings and integrates them within country-specific and cultural contexts. This graph supports a RAG pipeline in which relevant knowledge is retrieved from the graph and provided as contextual grounding to an LLM, enhancing the model's ability to interpret and generate outputs that are both emotionally accurate given the cultural context.

3 Methodolody

This section outlines the SLANG-GraphRAG process, focusing on constructing diverse and high-quality cross-cultural social media datasets for emotion classification. The detailed architecture is illustrated in Figure 1.

3.1 Data Collection and Pre-processing

3.1.1 Dataset

Our study examined culturally-loaded social media data by collecting 138,862 geo-tagged tweets via X's (formerly Twitter) API, spanning January 2010 to March 2024. Tweets were collected using a set of keywords and hashtags, including climate change, global warming, deforestation, extinction, climate action, nature, biodiversity, and conservation. To ensure balanced cross-cultural representation, we explicitly constrained data collection to obtain equal numbers of tweets from the UK and Nigeria, yielding 69,724 tweets per country. Our research enhances the understanding of informal discourse, specifically how emotional expressions are conveyed across diverse cultures and dialects, building on earlier studies from non-Anglocentric regions. This offers an ideal resource for testing RAG systems that need to account for cultural and emotional nuances in an informal setting.

3.1.2 Slang Dictionary Generation

We curated a comprehensive slang dictionary comprising approximately 240 terms, covering slang words, phrases, and their definitions in both UK and Nigerian English. Slang entries were collected from a combination of regionally grounded online forums and linguistic databases, ensuring coverage of culturally embedded and contemporary usage. UK slang sources included Tandem.net, Urban Dictionary, Smartcat.com, and Parade.com, while Nigerian slang was sourced from Zikoko.com, Naijalingo.com, BBC Pidgin, and Urban Dictionary. These sources were selected to capture both widely used and locally nuanced expressions from both regions.

Slang identification was conducted using a dictionary-based matching approach, in which tweets were scanned for occurrences of terms listed in the curated slang dictionary. Additional filtering steps were applied to distinguish slang usage from standard lexical usage, particularly for terms that commonly appear in both forms. Ambiguous entries were manually reviewed and excluded unless contextual cues clearly indicated informal or slang usage. For example, the Nigerian slang term "oyo" commonly denotes "on your own," but also refers to a Nigerian state, while the UK slang term "sick" can express approval or excitement but is frequently used to describe illness. Such cases were handled carefully to ensure that retained instances reflected informal language use rather than literal meaning. Applying this filtering process resulted in 2,576 slang-containing climate-related tweets, of which 2,003 originated from Nigeria and 573 from the UK.

3.1.3 Data Annotation Setup

From the 2,576 slang tweets, we selected a 20% test set (517 tweets: 397 Nigerian and 120 UK) for manual annotation to establish a gold-standard

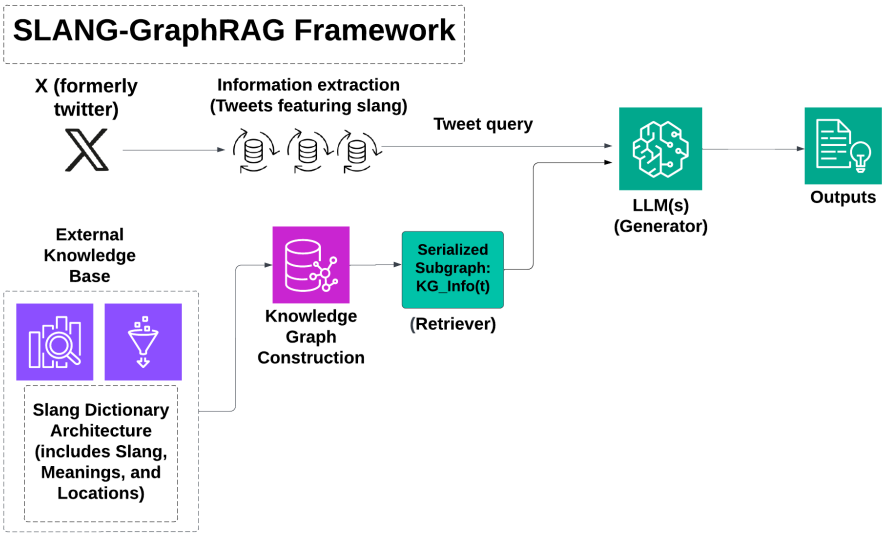


Figure 1: The SLANG-GraphRAG framework consists of three main stages: (1) construct a slang-informed knowledge graph (`KG_Info`), (2) retrieve relevant cultural context, and (3) generate emotion predictions using an LLM prompted on that context.

baseline. To ensure consistency, annotators were first shown a sample slang term with its usage in a tweet. Nine voluntary annotators per region, selected for their familiarity with local slang, labeled each tweet with one of the seven emotions (Joy, Sadness, Anger, Fear, Disgust, Neutral, Surprise).

3.2 Synthesizing SLANG-GraphRAG

In this section, we present SLANG-GraphRAG, a multi-stage pipeline for emotion prediction in informal conversations.

3.2.1 Retriever (Knowledge Graph Construction for Slang)

We construct a knowledge graph (KG) to represent the relationships between British and Nigerian slang words, their standard meanings, and associated locations or regional tags. We begin by scanning each tweet t for the existence of slang terms s drawn from our curated slang dictionary. If a term is detected, we pull its definition d (via a "means" edge), and when available, we retrieve its country of usage c (via a "used_in" edge) from a directed KG.

The formal definition of the graph is:

$$G = (V, E)$$

where

$$V = \{s\} \cup \{d\} \cup \{c\},$$
$$E = \{(s, d, \text{means}), (s, c, \text{used\_in})\}.$$

Here G is the graph, V its set of vertices (nodes), and E its set of directed edges. We serialize the resulting subgraph for a tweet t into a compact string, which we denote $\text{KG\_Info}(t)$. For example, `Abeg→'means'→Please; Abeg→'used_in'→Nigeria`

We implement a multi-layered retrieval process that enhances understanding by drawing on:

- D_1: slang dictionary data (custom-built Nigerian slang)
- D_2: region (slang-region)

Then, to enhance classification q (e.g., of a tweet featuring Nigerian slang), we retrieve relevant context as:

$$R = \text{Retrieve}(q, D_1, D_2, \ldots, D_k)$$

3.2.2 Generator (RAG Prompting)

We utilized one-shot prompting for emotion classification in tweets featuring slang from both regions, providing a single example to guide the model's understanding. After testing various strategies, one-shot prompting offered the best balance of clarity and effectiveness across seven emotion categories (Joy, Sadness, Anger, Fear, Disgust, Neutral, and Surprise) drawn from established emotion taxonomies such as Ekman's basic emotions (Ekman, 1992) and subsequent NLP adaptations (Hartmann, 2022). Combined with the knowledge graph, this

approach enables accurate classification of culturally nuanced, informal language. We construct the "one-shot" prompt by concatenating:

$$x' = \underbrace{\text{Example tweet + label}}_{E} ; \underbrace{t}_{\text{Tweet}} ; \underbrace{\text{KG\_Info}(t)}_{\text{Knowledge}} ; \text{ Answer:}$$

The Python-style template for the prompt is then:

```
kg_info = Abeg means Please; used_in Nigeria
prompt = f"""
You are an expert annotator for short tweets
    with slang. Classify the emotion using ONLY
    these labels: Joy, Anger, Sadness, Fear,
    Neutral, Disgust, Surprise. Use the provided
     knowledge graph information if relevant:{
    kg_info}
Example Tweet: "I'm so excited for the climate
    concert today! Abeg, who's coming with me?"
Emotion: Joy
Tweet: "{t}"
Answer:
"""
```

3.2.3 LLM Backends

We feed x' into several LLM backends including LLaMA-3.2-1B (AI@Meta, 2024), LLaMA-3.2-3B (AI@Meta, 2024), Qwen2-7B-Instruct (Alibaba, 2025), DeepSeek-llm-7b-chat (DeepseekAI, 2024). Each model produces a short continuation, and we select the first generated token as the predicted emotion label. Together, these components allow SLANG-GraphRAG to ground each tweet in structured, culturally grounded knowledge before emotion classification, improving the model's handling of informal, dialect-specific language. We prioritize models that are easily accessible to researchers to ensure reproducibility and transparency, excluding proprietary options like OpenAI's GPT variants. By selecting smaller, flexible open-source models, we strike a balance between efficiency and performance, supporting reliable and replicable research (Bender et al., 2021).

3.3 Evaluation Frameworks

To evaluate the performance of our SLANG-GraphRAG pipeline and existing LLMs and RAG framework, we employ two sets of metrics:

3.3.1 Standard classification metric

We employ standard evaluation metrics such as accuracy, precision, recall, and f1-score to measure classification performance of the models.

3.3.2 Probabilistic and Normalized Metrics

Standard evaluation metrics typically assume a single 'correct' label against which model predictions can be compared. However, emotion classification is a task for which human raters frequently disagree (Rizos and Schuller, 2019; Lotfian and Busso, 2019). Rather than treating this disagreement as noise or measurement error, we propose that it reflects the presence of multiple valid interpretations. Therefore, we assume each text is associated with a probability distribution over all emotion labels, and the task becomes to assess how likely the model's predictions are under this distribution.

We estimate the underlying distribution from human annotations and evaluate model outputs by computing the log-probability that the model's predicted labels were drawn from the estimated distribution. Specifically, for k possible emotion labels, we assume a uniform Dirichlet prior:

$$p \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_k),$$
$$\text{where all } \alpha_j = 1\text{e}{-10}, \text{ for } j = 1, \ldots, k$$

The small value of α avoids division by zero, while allowing human annotations to dominate the posterior. This formulation also permits future extensions using a hierarchical Bayesian model, where priors over α can be learned from a broader set of human annotations.

Given a particular text t_i, let $x^{(i)} = (x_1^{(i)}, \ldots, x_k^{(i)})$ represent the number of human annotators who assigned each label. We update our prior using these counts to obtain a posterior distribution:

$$p \mid x^{(i)} \sim \text{Dirichlet}(\alpha_1 + x_1^{(i)}, \ldots, \alpha_k + x_k^{(i)})$$

From this, we compute the posterior predictive probability for the model's predicted label $l_i \in \{1, \ldots, k\}$ for that text:

$$P(l_i \mid x^{(i)}) = \frac{\alpha_{l_i} + x_{l_i}^{(i)}}{\sum_{j=1}^{k} (\alpha_j + x_j^{(i)})}$$

We then define our evaluation metric as the total log-probability across all n data points:

$$\sum_{i=1}^{n} \log P(l_i \mid x^{(i)})$$

This metric rewards predictions that align well with the distribution of human judgments and allows for partial credit in cases where the model's prediction matches one of several plausible labels.

To make this metric interpretable, we compute multiple points of comparison that represent intuitive lower and upper bounds.

Lower bounds:

(a) **Random guessing:** Assign each text a label chosen uniformly at random. Repeating this process produces a distribution of log-probability scores representing chance performance.

(b) **Most common label:** Assign the most frequently chosen emotion across all texts to every data point. This models a classifier that always outputs the most common label.

Upper bounds:

(a) **Majority label:** Assign to each text the label chosen most frequently by its human annotators (i.e., the mode of $x^{(i)}$). This corresponds to a gold-standard approach that chooses the most likely label per instance, but weights instances by annotator agreement.

(b) **Random human choice:** For each text, select a label at random from among those assigned by human annotators. Repeating this process produces a distribution of log-probability scores representing the expected agreement of individual humans with the collective distribution.

Min-max standardization: To facilitate comparison, we compute a standardized score by normalizing the model's log-probability against the minimum and maximum expected values:

$$S = \frac{\sum_{i=1}^{n} \log P(l_i \mid x^{(i)}) - \log P_{\text{min}}}{\log P_{\text{max}} - \log P_{\text{min}}}$$

where $\log P_{\text{min}}$ is the mean of random guessing and $\log P_{\text{max}}$ is the majority label.

In future work, this framework can be extended to assess probabilistic model outputs by comparing the predicted distribution over emotion labels to the human-estimated distribution. This opens the door to evaluating not only the top prediction but the entire distribution of a model's uncertainty in a principled way.

4 Result and Discussion

4.1 Experimental Results

Across all LLMs, we use the same default sampling parameters: do_sample=False, temperature=1.0, and top_p=1.0. These settings ensure deterministic, consistent, and reproducible predictions across models. We extract the first generated token as the predicted emotion. All models were evaluated on the 20% human-annotated test set, and classification performance is reported in Table 1.

Experimental Models	Standard Classification Metrics			
	Nigeria Distribution		UK Distribution	
	Accuracy	F1_score	Accuracy	F1_score
Large Language Models				
LLaMA3.2-1B	0.35	0.38	0.47	0.49
LLaMA3.2-3B	0.45	0.43	**0.61**	**0.60**
Phi-3.5-mini-instruct	0.36	0.36	0.52	0.51
Qwen2-7B-Instruct	0.44	0.43	0.56	0.52
DeepSeek-llm-7b-chat	0.37	0.35	0.47	0.43
Standard RAG with slang dictionary				
LLaMA3.2-1B (Standard RAG)	0.41	0.40	0.51	0.43
LLaMA3.2-3B (Standard RAG)	0.54	0.51	0.57	0.53
Qwen2-7B-Instruct (Standard RAG)	0.34	0.35	0.46	0.45
Single-hop graph RAG Variant (only "has_meaning" in KG_Info)				
LLaMA3.2-1B (Single-hop)	0.35	0.39	0.46	0.48
LLaMA3.2-3B (Single-hop)	0.55	0.52	0.58	0.55
Qwen2-7B-Instruct (Single-hop)	0.47	0.44	0.49	0.45
Double-hop graph RAG Variant ("has_meaning" & "used_in" in KG_Info)				
LLaMA3.2-1B (Double-hop)	0.37	0.40	0.45	0.42
LLaMA3.2-3B (Double-hop)	**0.59**	**0.55**	0.57	0.54
Qwen2-7B-Instruct (Double-hop)	0.49	0.45	0.54	0.48

Table 1: Standard classification metrics by region for each model

The results highlight a clear disparity between Nigerian and UK data. Baseline models consistently performed better on UK tweets (e.g., LLaMA3.2-3B at 61% accuracy / 0.60 F1) than on Nigerian tweets (45% / 0.43 F1), reflecting the stronger representation of UK English in pretraining corpora. Performance on Nigerian data improved most when cultural context was introduced. With Double-hop GraphRAG, LLaMA3.2-3B increased to 59% accuracy / 0.55 F1, showing the benefit of linking slang with both meaning and regional usage. In contrast, the effect of cultural grounding was limited on UK slang tweets. For instance, LLaMA3.2-3B dropped slightly from 61% (baseline) to 57% accuracy with Double-hop GraphRAG. This reflects that LLMs are already relatively familiar with high-resource English corpora, so additional retrieval adds little or no significant value. These results demonstrate that multi-layered, culturally grounded graph retrieval methods substantially enhance performance on emotion classification tasks particularly in low-resource contexts. Human annotation reliability further reflects this challenge. Krippendorff's alpha was 0.19 for Nigerian tweets, compared to 0.41 for UK tweets, indicating far lower agreement among annotators in Nigeria. This suggests that slang-rich Nigerian tweet is more ambiguous and culturally dependent, making emotion classification difficult even for hu-

Experimental Models	Probabilistic & Normalized Metrics	
	log sum	Min–Max standardized log sum
Human annotations (Baseline)		
Log_max_emotion (Most Common Label)	-2438.31	0.57
Log_sums_fair (Majority Label)	-269.27	0.97
Large Language Models		
LLaMA3.2-1B	-2803.13	0.50
LLaMA3.2-3B	-2694.05	0.52
Phi-3.5-mini-instruct	-3102.46	0.44
Qwen2-7B-Instruct	-2605.43	0.54
DeepSeek-llm-7b-chat	-2962.74	0.47
Standard RAG with slang dictionary		
LLaMA3.2-1B (Standard RAG)	-2590.32	0.55
LLaMA3.2-3B (Standard RAG)	-1813.98	0.69
Qwen2-7B-Instruct (Standard RAG)	-2877.31	0.49
Single-hop graph RAG Variant (with only "means" in KG_Info)		
LLaMA3.2-1B (Single-hop graph RAG)	-2187.94	0.62
LLaMA3.2-3B (Single-hop graph RAG)	-1718.64	0.71
Qwen2-7B-Instruct (Single-hop graph RAG)	-2068.70	0.58
Double-hop graph RAG Variant (with "means" & "used_in"in KG_Info)		
LLaMA3.2-1B (Double-hop graph RAG)	-1770.46	0.71
LLaMA3.2-3B (Double-hop graph RAG)	**-1655.45**	**0.73**
Qwen2-7B-Instruct (Double-hop graph RAG)	-2427.57	0.65

Table 2: Probabilistic & Normalized Metrics of Experimental Models

mans. Such low agreement is typical of emotion annotation tasks (Rizos and Schuller, 2019; Sharma et al., 2019; Canales et al., 2022), underlining the need for evaluation metrics that account for label uncertainty.

In Table 2, we use the human-labeled set as a benchmark to compare model predictions, providing insight into where LLM-based reasoning aligns with or diverges from human judgment, particularly in slang-laden Nigerian tweets. The human baseline: Most Common Label = 0.57 and Majority label = 0.97 sets for comparison. Among the LLMs, Phi-3.5-mini-instruct records the lowest normalized score (0.44), while Qwen2-7B reaches the highest (0.54), both falling below the lower bound of choosing the most common label for all tweets. Standard RAG, which retrieves from a slang dictionary without graph structure, improves scores, raising LLaMA3.2-1B to 0.55 and LLaMA3.2-3B to 0.69. The Single-hop GraphRAG variant further enhances performance by explicitly linking slang terms to their definitions, boosting LLaMA3.2-1B to 0.62 and LLaMA3.2-3B to 0.71. Finally, the Double-hop GraphRAG achieves the strongest results: 0.71 for LLaMA3.2-1B and 0.73 for LLaMA3.2-3B, indicating that the combination of both definitions and contextual information about countries results in well-grounded predictions.

4.2 Impact of KG Retrieval and Cultural Sensitivity

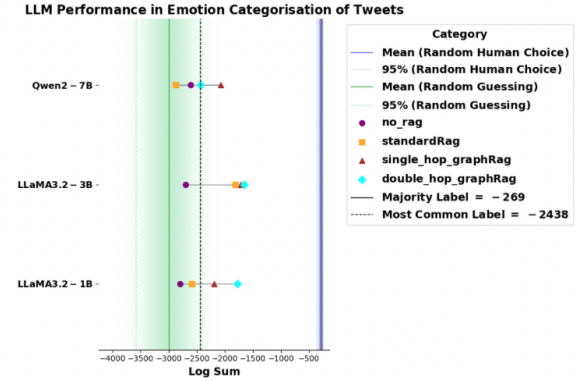
Our experiments, in Table 1 and Table 2, demonstrate the significant impact of incorporating structured, culturally sensitive KG in the RAG framework or Nigerian slang tweets. Simply adding slang terms through the Standard RAG approach provided some improvements; for example, LLaMA3.2-3B improved from 0.45 to 0.54 accuracy, with its normalized log-sum rising from 0.52 to 0.69. Also, the Single-hop GraphRAG method, which links slang terms to their definitions (via the "means" edges), produced stronger improvements. LLaMA3.2-3B achieved 0.55 accuracy and a normalized score of 0.71, indicating that incorporation of structured meaning through a knowledge graph facilitates the model's interpretation of slang and emotional expressions with greater reliability.

The most notable improvements, however, were seen in Figure 2a when we employed the Double-hop Graph RAG method. This approach not only connects slang terms to their definitions but also integrates country-specific contextual information. This double-layered approach provided the LLaMA-3.2-3B model with a richer understanding of slang in culturally specific contexts, leading to even higher performance: accuracy increased to 0.59, and the normalized log-sum increased to 0.73. These results underscore the importance of integrating both linguistic meaning and cultural context when working with informal, slang-featuring language, particularly in low-resource settings. This aligns with prior findings that highlight strong cultural and linguistic variation in non-Anglocentric varieties of English (Wuraola et al., 2023), as well as recent work showing that integrating culturally relevant knowledge improves model performance in low-resource, domain-specific NLP tasks (Ramadani and Doko, 2025; Ndimbo et al., 2025).

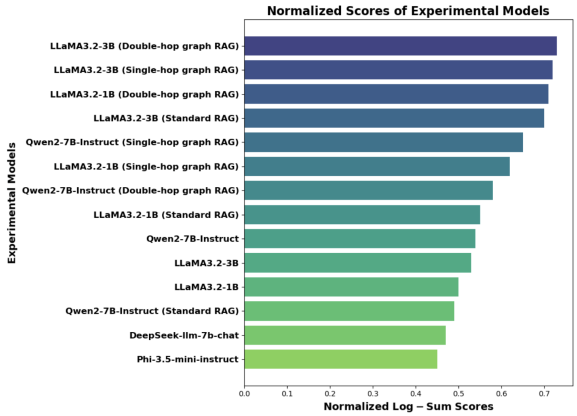
4.3 Evaluation and Human Alignment

Our results demonstrate that relying solely on traditional evaluation metrics like accuracy and F1-score provides an incomplete picture of model performance particularly for subjective tasks like emotion classification, where multiple interpretations can be valid. Our ambiguity-aware evaluation complements standard metrics by assessing how well model predictions align with the full distribution of human annotations, not just the majority vote.

For instance, in Table 1 LLaMA-3.2-3B im-

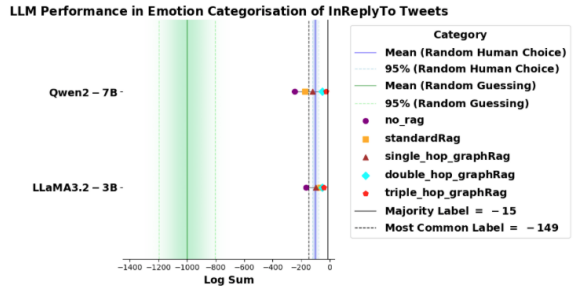


(a) LLM Performance in Emotion Classification of Tweets

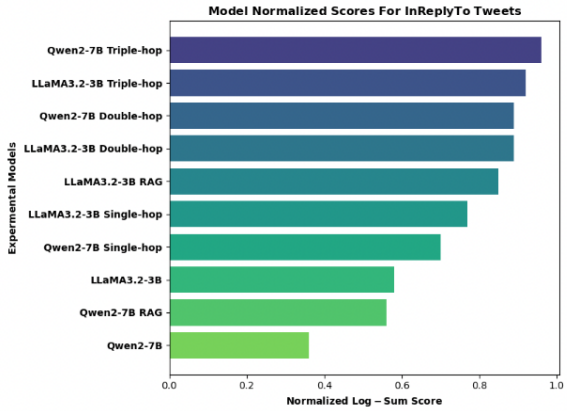


(b) Confidence of Experimental Models

Figure 2: Model Performance and Confidence Across Experimental Models



(a) LLM Performance in Emotion Classification of Prior Conversation



(b) Model Confidence of Prior Conversation

Figure 3: Experimental Result of Prior Conversation

proves from 0.45 to 0.55 in accuracy with single-hop graph context, but more notably, its confidence score increases from 0.52 to 0.71 (Table 2). This suggests not only that the model is producing more correct outputs, but that its predictions better reflect the range of human interpretations. Additionally, our graphRAG methods, especially double-hop retrieval, produce more stable and better-aligned predictions. Table 2 shows LLaMA-3.2-3B reaches a normalized log score of 0.73, a significant gain close to the level of consistency seen in individual human annotators (Majority Label = 0.97). Ultimately, ambiguity-aware evaluation offers a clearer measure of model–human agreement, especially in culturally complex emotion recognition tasks where multiple interpretations are expected.

4.4 Impact of Conversational Signals

To better capture conversational nuances in tweets, especially regarding how a user's reply may shift emotional tone, we are introducing a third retrieval signal (D_3) to SLANG-GraphRAG. For each tweet t, we check if it has "inReplyToTweet" column and introduced the replies into our KG_Info as "text_parent". In our dataset, a total of 26 slang-rich climate tweets are replies. We repeat our experiment using the best-performing LLMs (LLaMA3.2-3B and Qwen2-7B-Instruct) and report in Table 3.

Table 3 and Figure 3 illustrate how increasingly structured and context-rich retrieval strategies affect model performance, as measured by log-probability scores. In Table 3, the human-derived scores: Most Common Label = 0.62 and Majority Label = 0.97 set the performance benchmarks for emotion classification. Initial results from zero-shot LLMs such as LLaMA-3.2-3B and Qwen2-7B-Instruct show relatively low normalized log-sum (0.57 and 0.36, respectively), suggesting difficulty in interpreting informal, culturally embedded language without external cues. Standard RAG, Single-hop graph RAG, and Double-hop graph RAG with slang context gradually increases the

	Probabilistic & Normalized Metrics	
Experimental Models	**Log sum**	**Min–Max standardized log sum**
Human annotations (Baseline)		
Log_max_emotion (Most Common Label)	-149.88	0.62
Log_sums_fair (Majority Label)	-15.26	0.97
Large Language Models		
LLaMA3.2-3B	-167.83	0.57
Qwen2-7B-Instruct	-245.94	0.36
Standard RAG with slang dictionary		
LLaMA3.2-3B (Standard RAG)	-69.79	0.84
Qwen2-7B-Instruct (Standard RAG)	-174.67	0.56
Single-hop graph RAG (with only "means" in KG_Info))		
LLaMA3.2-3B (1-hop)	-97.36	0.77
Qwen2-7B-Instruct (1-hop)	-122.65	0.70
Double-hop graph RAG (with "means" & "used_in" in KG_Info)		
LLaMA3.2-3B (2-hop)	-53.16	0.89
Qwen2-7B-Instruct (2-hop)	-52.29	0.89
Triple-hop graph RAG (with "means", "used_in", "text_parent" in KG_Info))		
LLaMA3.2-3B (3-hop)	-42.45	0.92
Qwen2-7B-Instruct (3-hop)	**-26.90**	**0.96**

Table 3: Probabilistic & Normalized Metrics of Experimental Models of Prior Conversation

models' performance with Double-hop graphRAG leading. The most substantial gains appear in Figure 3a with the Triple-hop graph RAG setup, which includes slang definitions, region-specific, and prior conversations from parent tweets. Here, LLaMA and Qwen models achieve normalized log-sum of 0.92 and 0.96 respectively, almost reaching the human-annotated responses. Together, these findings reinforce the need for culturally and conversationally aware retrieval strategies.

5 Conclusion

This work shows that integrating culturally grounded knowledge into LLM prompts significantly improves emotion recognition in non-Anglocentric social media text. Our SLANG-GraphRAG framework, which retrieves structured slang definitions and usage contexts from a tailored KG, outperforms standard one-shot prompting and retrieval baselines. Notably, the double-hop Graph-RAG with LLaMA-3.2-3B boosts accuracy by 31.1% (from 0.45 to 0.59 accuracy) and raises normalized log-sum from 0.52 to 0.73. By combining multi-layered retrieval with a probabilistic evaluation metric, we capture both performance gains and model calibration aligned with human judgment. While conversational context can enhance prompts, our results emphasize the importance of cultural sensitivity where explicit, domain-specific knowledge enables more accurate interpretation of slang-rich, regional language. This approach paves the way for more context-aware NLP across diverse linguistic communities.

Although SLANG-GraphRAG advances culturally sensitive emotion modeling, opportunities remain to further enhance it. Automating slang discovery to update the KG, supporting multi-token and hierarchical emotion labels for granular sentiment analysis, and expanding to additional dialects, code-mixed text, and low-resource languages.

6 Limitations

While SLANG-GraphRAG demonstrates that integrating culturally grounded knowledge into LLM prompts can significantly improve emotional understanding of non-Anglocentric social media text, it does have some limitations. Specifically, the system's reliance on a manually curated slang dictionary restricts the range of slang terms it can handle. Emerging or unlisted slang may not be captured, and creating such resources for every new dialect or language requires substantial human effort. Additionally, we focused on small, accessible open-source models to ensure academic reproducibility and reduce resource demands, however, future research could explore the performance of larger-scale models.

Furthermore, our study is limited to specific demographics (Nigerian and UK) social media text to capture both a low-resource, non-Anglocentric setting and a high-resource English baseline. Attempting to replicate this work across multiple non-Anglocentric cultures would require deep cultural expertise in each case, as slang and emotional expression are highly context-dependent. Due to these resource and cultural constraints, we limited our analysis to one underrepresented region (Nigeria) alongside a well-resourced reference point (UK). Future work could expand our framework to additional dialects, code-mixed text, and underrepresented communities, thereby enhancing generalizability while accounting for the unique cultural nuances of each linguistic setting.

7 Ethics Statement

This study was conducted in full compliance with the ACL Ethics Policy to ensure ethical integrity and responsible research practices. Data collection was limited exclusively to publicly accessible tweets, with rigorous anonymization procedures applied to protect individual privacy and confidentiality. Additionally, we avoid reinforcing harmful biases or cultural stereotypes, approaching the research with respect for diverse cultural norms and values. The work makes use of suitable compu-

tational and statistical techniques, and we openly communicated our results to the larger scientific community. Our commitment to ethical standards underpins every stage of this research.

Acknowledgments The authors express gratitude to the Center of Excellence for Data Science, Artificial Intelligence and Modeling (DAIM) and the Language and Data Research (LDR) group for generously funding and enabling this research. We acknowledge the VIPER high-performance computing facility of the University of Hull and its support team.

References

Syed Rameel Ahmad. 2024. Enhancing Multilingual Information Retrieval in Mixed Human Resources Environments: A RAG Model Implementation for Multicultural Enterprise. *arXiv preprint*. ArXiv:2401.01511 [cs].

AI@Meta. 2024. Llama 3 Model Card. Original-date: 2024-03-15T17:57:00Z.

Alibaba. 2025. Qwen/Qwen2-7B-Instruct · Hugging Face.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating Cultural Alignment of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Mitha Alshammary, Md Nahiyan Uddin, and Latifur Khan. 2024. RFPG: Question-Answering from Low-Resource Language (Arabic) Texts using Factually Aware RAG. *2024 IEEE 10th International Conference on Collaboration and Internet Computing (CIC)*, pages 107–116. Conference Name: 2024 IEEE 10th International Conference on Collaboration and Internet Computing (CIC) ISBN: 9798350386707 Place: Washington, DC, USA Publisher: IEEE.

Maha Jarallah Althobaiti. 2023. Arabic Emotion Recognition in Low-Resource Settings: A Novel Diverse Model Stacking Ensemble with Self-Training. *Applied Sciences*, 13(23):12772. Number: 23 Publisher: Multidisciplinary Digital Publishing Institute.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada. ACM.

Yucheng Cai, Si Chen, Yuxuan Wu, Yi Huang, Junlan Feng, and Zhijian Ou. 2024. The 2nd Futuredial Challenge: Dialog Systems With Retrieval Augmented Generation (Futuredial-RAG). In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1091–1098.

Lea Canales, Walter Daelemans, Ester Boldrini, and Patricio Martínez-Barco. 2022. EmoLabel: Semi-Automatic Methodology for Emotion Annotation of Social Media Text. *IEEE Transactions on Affective Computing*, 13(2):579–591. Conference Name: IEEE Transactions on Affective Computing.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762. Number: 16.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

DeepseekAI. 2024. deepseek-ai/deepseek-llm-7b-chat · Hugging Face.

Jialin Dong, Bahare Fatemi, Bryan Perozzi, Lin F. Yang, and Anton Tsitsulin. 2024. Don't Forget to Connect! Improving RAG with Graph-based Reranking. *arXiv preprint*. ArXiv:2405.18414 [cs].

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint*. ArXiv:2404.16130 [cs].

Paul Ekman. 1992. Are There Basic Emotions? *Psychological review*, 99:550–3.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint*. ArXiv:2312.10997 [cs].

Haoyu Han, Harry Shomer, Yu Wang, Yongjia Lei, Kai Guo, Zhigang Hua, Bo Long, Hui Liu, and Jiliang Tang. 2025. RAG vs. GraphRAG: A Systematic Evaluation and Key Insights. *arXiv preprint*. ArXiv:2502.11371 [cs].

Jochen Hartmann. 2022. Emotion English DistilRoBERTa-base.

Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2024. GRAG: Graph Retrieval-Augmented Generation. *arXiv preprint*. ArXiv:2405.16506 [cs].

Atia Shahnaz Ipa, Mohammad Abu Tareq Rony, and Mohammad Shariful Islam. 2025. Empowering Low-Resource Languages: TraSe Architecture for Enhanced Retrieval-Augmented Generation in Bangla. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 8–15, Albuquerque, New Mexico. Association for Computational Linguistics.

Sahil Kamath, Vaishnavi Padiya, Sonia D'Silva, Nilesh Patil, and Meera Narvekar. 2024. TeenSenti - A novel approach for sentiment analysis of short words and slangs. In *2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE)*, pages 1–8.

Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A Causal Analysis of Semantic Change and Frequency Dynamics in Slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1422–1442, Dublin, Ireland. Association for Computational Linguistics.

Quinn Leng, Jacob Portes, Sam Havens, Matei Zaharia, and Michael Carbin. 2024. Long Context RAG Performance of Large Language Models. *arXiv preprint*. ArXiv:2411.03538 [cs].

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Mufei Li, Siqi Miao, and Pan Li. 2025. Simple Is Effective: The Roles of Graphs and Large Language Models in Knowledge-Graph-Based Retrieval-Augmented Generation. *arXiv preprint*. ArXiv:2410.20724 [cs].

Xiaoqian Li, Ercong Nie, and Sheng Liang. 2023. Crosslingual Retrieval Augmented In-context Learning for Bangla. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 136–151, Singapore. Association for Computational Linguistics.

Chang Liu, Loc Hoang, Andrew Stolman, and Bo Wu. 2024. HiTA: A RAG-Based Educational Platform that Centers Educators in the Instructional Loop. In *Artificial Intelligence in Education*, pages 405–412, Cham. Springer Nature Switzerland.

Reza Lotfian and Carlos Busso. 2019. Curriculum Learning for Speech Emotion Recognition From Crowdsourced Labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4):815–826.

Shuangjia Lu and Erdal Cosgun. 2025. Boosting GPT models for genomics analysis: generating trusted genetic variant annotations and interpretations through RAG and Fine-tuning. *Bioinformatics Advances*, 5(1):vbaf019.

Edmund V. Ndimbo, Qin Luo, Gimo C. Fernando, Xu Yang, and Bang Wang. 2025. Leveraging Retrieval-Augmented Generation for Swahili Language Conversation Systems. *Applied Sciences*, 15(2):524. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Anh-Khoi Ngo-Ho, Khuong-Duy Vo, and Anh-Khoa Ngo-Ho. 2025. Function-Based Rag for Vietnamese Speech Comprehension in Service Robotics. In *2025 5th Asia Conference on Information Engineering (ACIE)*, pages 100–104.

Ke Ni and William Yang Wang. 2017. Learning to Explain Non-Standard English Words and Phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, and Olalekan Akinsande. 2020. Semantic Enrichment of Nigerian Pidgin English for Contextual Sentiment Classification. *arXiv preprint*. ArXiv:2003.12450 [cs].

Leotrim Ramadani and Fisnik Doko. 2025. Exploring RAG Solutions for a Specific Language: Albanian. *European Journal of Information Technologies and Computer Science*, 5(1):26–31. Number: 1.

Georgios Rizos and Bjorn Schuller. 2019. Modelling Sample Informativeness for Deep Affective Computing. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3482–3486, Brighton, United Kingdom. IEEE.

Bhargav Shandilya and Alexis Palmer. 2025. Boosting the Capabilities of Compact Models in Low-Data Contexts with Large Language Models and Retrieval-Augmented Generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7470–7483, Abu Dhabi, UAE. Association for Computational Linguistics.

Karan Sharma, Marius Wagner, Claudio Castellini, Egon L. van den Broek, Freek Stulp, and Friedhelm Schwenker. 2019. A functional data analysis approach for continuous 2-D emotion annotations. *Web Intelligence*, 17:41–52.

Peng Shu, Junhao Chen, Zhengliang Liu, Hui Wang, Zihao Wu, Tianyang Zhong, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Yifan Zhou, Constance Owl, Xiaoming Zhai, Ninghao Liu, Claudio Saunt, and Tianming Liu. 2024. Transcending Language Boundaries: Harnessing LLMs for Low-Resource Language Translation. *arXiv preprint*. ArXiv:2411.11295 [cs].

Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. Toward Informal Language Processing: Knowledge of Slang in Large Language Models. *arXiv preprint*.

Elizaveta Vologina, Anastasiia Matveeva, Olesia Makhnytkina, Yuri Matveev, and Nursaule Burambayeva. 2024. RAG and Few-Shot Prompting in Emotional Text Generation. In *Speech and Computer*, pages 43–53, Cham. Springer Nature Switzerland.

Yuwei Wan, Zheyuan Chen, Ying Liu, Chong Chen, and Michael Packianather. 2025. Empowering LLMs by hybrid retrieval-augmented generation for domain-centric Q&A in smart manufacturing. *Advanced Engineering Informatics*, 65. Publisher: Elsevier.

Mingyang Wang, Alisa Stoll, Lukas Lange, Heike Adel, Hinrich Schütze, and Jannik Strötgen. 2025. Bring Your Own Knowledge: A Survey of Methods for LLM Knowledge Expansion. *arXiv preprint*. ArXiv:2502.12598 [cs].

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2024. Retrieval-Augmented Multilingual Knowledge Editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 335–354, Bangkok, Thailand. Association for Computational Linguistics.

Steven Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. 2020. Urban Dictionary Embeddings for Slang NLP Applications. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4764–4773, Marseille, France. European Language Resources Association.

Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. 2023. Linguistic Pattern Analysis in the Climate Change-Related Tweets from UK and Nigeria. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 90–97, Gothenburg, Sweden. Association for Computational Linguistics.

Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. 2024. Understanding Slang with LLMs: Modelling Cross-Cultural Nuances through Paraphrasing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15525–15531, Miami, Florida, USA. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv preprint*. ArXiv:2305.10601 [cs].

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, and Tianming Liu. 2024. Opportunities and Challenges of Large Language Models for Low-Resource Languages in Humanities Research. *arXiv preprint*. Publisher: arXiv Version Number: 2.

A Supplementary Materials

A.1 Data Sources

A.1.1 Dataset Access

Due to platform restrictions, full tweet text cannot be redistributed. To support transparency and reproducibility, we release Tweet IDs and associated annotations, including timestamps, country attribution, and slang indicators. The dataset is publicly available on Zenodo:

https://doi.org/10.5281/zenodo.18102351

Researchers can rehydrate the tweets in compliance with platform terms of service.

A.1.2 Slang Dictionary Sources

The slang dictionary was curated from region-specific linguistic and media resources to capture informal and culturally grounded language use. Sources include:

UK:

- Tandem.net: https://www.tandem.net/blog/british-slang-words
- Urban Dictionary: https://www.urbandictionary.com
- Smartcat.com: https://www.smartcat.com/blog/100-british-slang-words-and-expressions-to-knock-your-socks-off/
- Parade.com: https://parade.com/1293790/marynliles/british-slang-words/

Nigeria:

- Zikoko.com: https://www.zikoko.com/pop/nigerian-slangs-and-their-meanings-the-2024-guide/
- Naijalingo.com: http://naijalingo.com/words/a/alphabet
- BBC Pidgin: https://www.bbc.com/pidgin/articles/cxwk8ldp7mno
- Urban Dictionary: https://www.urbandictionary.com

A.2 Annotation

A.2.1 Annotation Guidelines and Definitions

Tweets containing slang or informal language were annotated for emotional content. Annotators were instructed to identify the *single dominant emotion*

expressed in each tweet. Nine voluntary annotators per region, selected for their familiarity with local slang were instructed to identify the *single dominant emotion* expressed in each tweet. Each tweet was labeled using one of the following emotion categories: *anger, fear, disgust, joy, sadness, surprise,* and *neutral*.

Annotators were asked to select the emotion that best reflects the overall tone and communicative intent of the tweet, even when multiple emotions were present.

Example

> **Tweet:** "I'm so excited for the climate concert today! Abeg, who's coming with me?"
> **Emotion:** *Joy*

A.2.2 Emotion Definitions

Each tweet was assigned one dominant emotion reflecting its primary affective signal. The emotion categories are defined as follows:

- **Anger**: Expressions of frustration, blame, outrage, or hostility.
- **Fear**: Expressions of concern, anxiety, uncertainty, or apprehension.
- **Disgust**: Language conveying strong disapproval, contempt, or moral revulsion.
- **Joy**: Expressions of optimism, relief, approval, or positive sentiment.
- **Sadness**: Expressions of loss, grief, resignation, or despair.
- **Surprise**: Reactions indicating shock or unexpectedness, often in response to sudden events.
- **Neutral**: Informational or descriptive content without a clear emotional signal, including factual reporting or explanatory statements.

A.2.3 Final Label Determination

To obtain a single gold emotion label per tweet, we applied majority voting across annotators. The emotion selected by the largest number of annotators was assigned as the final label. In cases where no clear majority was reached, the tweet was excluded from the annotated subset to ensure label reliability.

A.3 Large Language Models Used

We evaluated SLANG-GraphRAG using a range of LLMs to assess robustness across model variants and sizes. The following models were used in our experiments:

- **LLaMA 3.2 (1B, 3B)**
  `https://ai.meta.com/llama/`
- **Qwen-7B**
  `https://huggingface.co/Qwen`
- **DeepSeek-7B**
  `https://huggingface.co/deepseek-ai`
- **Phi-3.5**
  `https://huggingface.co/microsoft`

All models were accessed through their publicly available checkpoints and used in inference-only mode.

A.4 Experimental Settings

Across all LLMs, we applied the same inference settings to ensure comparability. Generation was performed with `do_sample=False`, `temperature=1.0`, and `top_p=1.0`. Using deterministic decoding (`do_sample=False`) ensures stable and reproducible outputs across runs. The temperature and nucleus sampling parameters were kept at their default values to avoid introducing additional variability across models. For classification tasks, we extract the first generated label token and use it as the predicted emotion.