Understanding Slang with LLMs: Modelling Cross-Cultural Nuances through Paraphrasing

Ifeoluwa Wuraola[1], Nina Dethlefs[1], Daniel Marciniak[2]
[1] School of Computer Science, University of Hull, UK,
[2] School of Criminology, Sociology and Policing, University of Hull, UK,
{i.a.wuraola-2021, n.dethlefs, d.f.marciniak}@hull.ac.uk

Abstract

In the realm of social media discourse, the integration of slang enriches communication, reflecting the sociocultural identities of users. This study investigates the capability of large language models (LLMs) to paraphrase slang within climate-related tweets from Nigeria and the UK, with a focus on identifying emotional nuances. Using DistilRoBERTa as the baseline model, we observe its limited comprehension of slang. To improve cross-cultural understanding, we gauge the effectiveness of leading LLMs: ChatGPT 4, Gemini, and LLaMA3 in slang paraphrasing. While ChatGPT 4 and Gemini demonstrate comparable effectiveness in slang paraphrasing, LLaMA3 shows less coverage, with all LLMs exhibiting limitations in coverage, especially of Nigerian slang. Our findings underscore the necessity for culturally-sensitive LLM development in emotion classification, particularly in non-anglocentric regions.

1 Introduction

In the age of social media, platforms like X (formerly Twitter) have become a vital medium for public discourse, where users express a wide array of views and emotions on various topics (Geronikolou et al., 2021; Loureiro and Alló, 2020; Wang et al., 2016). However, sociocultural identities including regional background, gender, age, and sub-cultural affiliations have a big impact on communication styles. People often blend formal and informal language, incorporating specific dialects or slang into their discourse. Discourse from non-Anglocentric countries may thus contain cultural references and idioms that are not easily understood by outsiders. For instance, Nigerian tweets may utilize Pidgin English to convey emotions, such as the phrase "I defy happy no be small" meaning "I am very happy".

text[[135, 859, 481, 923], [504, 252, 851, 499]] Emotion classification is a key task in sentiment analysis. Despite LLMs' impressive capabilities in various linguistic tasks, they often encounter challenges in accurately capturing cultural nuances like emotions, resulting in inaccuracies, particularly in diverse settings (Mao et al., 2023). In this paper we focus on LLMs' knowledge of slang and how state-of-the-art models might misinterpret or overlook emotions in tweets containing slang across different varieties of English. We propose a novel approach to integrating detailed slang representations into LLMs. Leveraging generative models such as OpenAI's ChatGPT 4 (OpenAI, 2024), Google's Gemini (GoogleAI, 2024), and META's LLaMA3 (AI@Meta, 2024), we systematically investigate how paraphrased slang influences emotional expressions in tweets from diverse cultures, focusing specifically on Nigeria and the UK. We make the following contributions:

We highlight shortcomings in pre-trained LLMs in identifying emotions in social media discourse featuring slang, particularly in non-Anglocentric varieties of English. We provide a comprehensive comparison of leading LLMs in understanding and paraphrasing slang, pointing to ways of reducing bias.

Our study highlights the need to model slang in reducing biases within LLMs, especially in regions with diverse linguistic backgrounds. Our research demonstrates the cultural insensitivity of LLMs for emotion classification in tweets from Nigeria and the United Kingdom (UK). By incorporating Nigerian perspectives, we address a critical gap in understanding cultural nuances and linguistic expressions in underrepresented groups.

2 Related Works

2.1 Cross-cultural performance of LLMs

Recent research has placed an increasing emphasis on addressing the complex interplay between cross-cultural context and bias mitigation in LLMs. Hershcovich et al. (2022) argue that current LLMs do not adequately model the intricate relationships between linguistic constructions and sociocultural viewpoints, values and common ground. Multiple studies have found that while LLMs perform well at standard English tasks, they are much less successful at modelling non-standard varieties of English, including African American English (Deas et al., 2023), non-Anglocentric varieties of English (Wuraola et al., 2023), Pidgin (Chang et al., 2022), or examples of code-switching Zhang et al. (2023).

Similarly, multi-lingual LLMs have been found to be much less reliable in practice than their English counterparts, including factual information systems (Fierro and Sogaard, 2022), emotion and sentiment classifica-

tion (Muhammad et al., 2023). Machine translation can affect the reliability of cross-cultural analyses (Zhang et al., 2023), particularly when LLMs transfer stereotypes between languages. Low-resource languages are especially susceptible to these leakages compared to dominant languages Cao et al. (2024). Dodge et al. (2021) demonstrate a bias towards US data in NLP resources and find that when data gets removed (e.g. toxicity, slurs, obscenity, etc.), this disproportionately affects data relating to minority groups.

2.2 Modelling Slang with LLMs

In this paper, we focus on the effect of slang on the task of emotion classification, specifically comparing British and Nigerian English. This links with previous studies that have explored cross-cultural context in slang analysis. Lin et al. (2018) introduce SocVec, which aims to compute cross-cultural differences in understanding slang terms across languages. The method is evaluated on two tasks focused on mining cross-cultural differences in named entities and slang. Similarly, Sun et al. (2024) use LLMs to detect slang and attribute regional and historical context. Despite GPT-4's high performance in zero-shot settings, the study reveals that smaller, fine-tuned BERT models achieve comparable results. Both studies underscore the significance of regional and cultural contexts in understanding slang.

text[[130, 764, 475, 926], [497, 76, 842, 304]] In a similar vein, Sun et al. (2021) introduced a computational framework for slang generation that incorporates syntactic and contextual knowledge. The framework leverages probabilistic inference and neural contrastive learning and outperforms existing language models in accurately predicting historical slang emergence from the 1960s to 2000s. Pei et al. (2019) compare classifiers for slang detection and highlight the syntactic shift of words as a key feature of slang. Seki and Liu (2022) enhanced LLMs for Chinese slang comprehension contrasting LLM performance with a custom Punchline Entity Recognition (PER) system, integrating phonetic matching. Also, Sultan (2023) classify emotions in tweets containing slang based on WordNet for synonymous phrase generation and a CNN for classification. They show a significant improvement in emotion classification for slang-filled social media texts. Furthermore, Rohn (2024) detect internet slang based on a hierarchical multi-task BERT model, using two-layer annotation and word embeddings. The model excelled in identifying subcategories of internet slang, demonstrating the effectiveness of two-layer annotation.

3 Methodology

3.1 Dataset

Our study explores climate-related tweets from Twitter (now X) sourced via the API and spanning a time frame of January 2010 to March 2024. Our data collection focused on keywords and hashtags related to climate change, global warming, and conservation, see Wuraola et al. (2023) for details. We balanced our data to make up equal proportions of tweets originating from the UK and Nigeria, via geo-tags, which led to a corpus of 138,862 tweets for analysis. The motivation for studying climate change tweets lies in the high volume and emotional intensity of discussions surrounding this topic on social media. Climate change is a global issue that elicits strong reactions and diverse linguistic expressions, including slang. Additionally, the two countries we focused on, the UK and Nigeria, are likely affected by climate change in different ways, making this an interesting domain to study. Any misinterpretations from an LLM could lead to a misrepresentation of views, which underscores the importance of accurately understanding the emotional content conveyed through slang.

3.2 Slang Dictionary Generation

In order to assess LLMs' ability to identify emotions in discourse containing slang, we initially evaluate their ability to paraphrase slang in UK and Nigerian English. For this purpose, we curated a comprehensive slang dictionary, consisting of about 240 unique slang terms and their meanings. These terms were sourced from a variety of channels, ensuring a diverse representation of contemporary slang that serves as our gold standard for paraphrasing. Specifically, we targeted online forums and linguistic databases relevant to each region, for example, Naijalingo.com for Nigerian slang and Tandem.net for UK slang. See Table 1 for details.

Table 1: Slang sources on the web

Country	Online slang sources
UK	Tandem.net, Urban dictionary, Smartcat.com, Parade.c
Nigeria	Zikoko.com, Naijalingo.com, BBC pidgin.com, Urban d

We compared the ability to generate concise paraphrases for slang terms of OpenAI's ChatGPT-4 (OpenAI, 2024), Google's Gemini (GoogleAI, 2024), and META's LLaMA3 8B (AI@Meta, 2024). ChatGPT and Gemini paraphrased all curated slang, whereas LLaMA3 was unable to provide a paraphrase for 22% of the slang. The paraphrases in Table 2 were generated using a zero-shot approach, where the LLMs were prompted to provide paraphrases for slang terms with regional specifications. For instance, we instructed the models with prompts like, paraphrase this Nigerian slang wahala' and paraphrase this UK slang 'mate'. Additionally, in Table 2, we assess the correctness of slang paraphrases against their dictionary definitions. The correctness of these paraphrases was evaluated through a manual review, where we compared the slang paraphrases to their meanings from the online sources.

LLaMA3 shows the lowest accuracy here, suggesting that the model may exhibit more bias in its slang knowledge from specific cultural contexts compared to Gemini and GPT models. This observation is further supported by employing Cohen's Kappa score, which demonstrates an agreement of 0.74 between ChatGPT

and Gemini for Nigerian tweets, the highest among all models. This metric solidifies the notion that ChatGPT and Gemini yield very similar effects, providing substantial evidence for their comparability.

Table 2: Correctly paraphrased tweets across LLMs.

	Nigeria	
	Incorrect paraphrases (%)	Correct paraphrases (%)
ChatGPT	8	92
Gemini	19	81
LLaMA3	45	55
	UK	
	Incorrect paraphrases (%)	Correct paraphrases (%)
ChatGPT	2	98
Gemini	4	96
LLaMA3	24	76

We used our four dictionary resources (i.e. manually curated, ChatGPT-4, Gemini and LLaMA3) to detect tweets containing one or more slang words or phrases and replaced them with their paraphrases. For this task, we employed a direct identification approach using our curated dictionary corpus to recognize and extract tweets containing slang terms from climate-related content. This process yielded a total of 2,845 tweets containing slang, with 592 originating from the UK and 2,253 from Nigeria. This confirms earlier research that exposed the linguistic variety in African English (Wuraola et al., 2023; Muhammad et al., 2023; Chang et al., 2022).

3.3 Emotion Labelling

We employ DistilRoBERTa (Hartmann, 2022) to label the emotions in our tweet dataset. This model features 6 transformer layers, a hidden size of 768 dimensions, and 12 attention heads, enabling it to effectively understand contextual information. During pre-training, DistilRoBERTa employs advanced feature extraction techniques to identify emotions, producing output vectors that represent seven distinct emotions (joy, sadness, anger, surprise, disgust, fear, and neutral). We perform this task twice: first on the original tweets containing slang, and then on the paraphrased versions. DistilRoBERTa labels were compared against ratings from five independent human raters. The raters achieved an agreement score of 0.30 with DistilRoBERTa and an agreement score of 0.41 among themselves. To further clarify our results, we manually examined the raters' labels and discovered that around 68% of them assigned two or more negative emotions to the same tweet. While individual raters may have chosen different specific labels, there was a general consensus on the overall emotional tone being negative. This indicates the complexity and nuanced nature of emotional expressions, underscoring the challenges in achieving consistent emotion identification (Sharma et al., 2019; Schoene et al., 2020; Canales et al., 2022).

4 Results and Discussion

In this section, we aim to determine the effect that slang, and understanding its correct meaning, has on emotion classification in tweets. To this end, we present two sets of results: (1) the emotion distribution in UK and Nigerian English in the original climate tweets in Section 3.1, and (2) the percentage changes in emotion distribution with each set of LLM-generated paraphrases in Section 3.2.

Figure 1 illustrates percentage changes between original and paraphrased tweets for each of the LLMs. It highlights distinct changes in emotion expression between Nigerian and UK tweets when paraphrased with different models. Nigerian tweets exhibit increased fear with ChatGPT (10.36%) , while UK tweets show decreased fear with LLaMA3 (−11.31%) and Gemini (−13.1%) . Anger decreases across both countries, notably in UK tweets paraphrased with ChatGPT (−61.54%) . Paraphrased tweets often show heightened joy, especially in Nigerian tweets paraphrased with LLaMA3 (148.48%) . Additionally, both countries experience reduced neutral emotions post-paraphrasing, indicating a shift towards more polarised language, or just highlighting that LLMs find it harder to discern emotions from slang.

figure1.png

Figure 1: Percentage Change in Emotions of Climate Tweets: Comparing Original and Paraphrased Versions from Nigeria and the UK.

Note: IMAGE NOT PROVIDED. Placeholder for Figure 1.

Table 3 compares emotion classification before and after paraphrasing. For instance, DistilRoberta initially classified 550 Nigerian tweets as expressing fear, increasing to 687 after paraphrasing with ChatGPT-4. These are notable shifts (both positively and negatively), indicating the baseline model's limited proficiency in emotion classification in the presence of slang. Inspecting the data, we find that Nigerian tweets featur-

ing the slang "wahala" are often misclassified as neutral. However, when paraphrased as "trouble" or "problem," the emotion changes to fear. For example, the Nigerian tweet "imagine that climate change switches everything and then it begins to snow in Nigeria wahala go dey oo", was paraphrased as, "imagine that climate change switches everything and then it begins to snow in Nigeria there will be trouble". This observation aligns with prior research emphasizing the importance of incorporating external context to enhance the LLMs' comprehension of social media data (Adedamola et al., 2015; Sultan, 2023).

text[[128, 894, 472, 926], [496, 75, 842, 271]] Overall, the effects of paraphrasing slang show significant variation between countries and models. Nigerian tweets typically demonstrate more pronounced emotional shifts across all models, with ChatGPT often amplifying emotions. In contrast, UK tweets exhibit more subtle changes, with LLaMA3, Gemini, and ChatGPT each impacting emotions differently. This indicates the significant influence of both cultural context and model-specific behaviour on emotion extraction. These variations may stem from inherent biases in LLMs towards underrepresented dialects, as evidenced by previous studies (Narayanan Venkit et al., 2023; Sun et al., 2019; Chuang et al., 2021).

Table 4 shows examples of paraphrased slang across English variants, indicating that ChatGPT and Gemini generally offer accurate paraphrases close to the gold standard. However, LLaMA3 falls behind in paraphrasing slang, implying a cultural bias compared to the other models. Table 2 supports this with LLaMA3 incorrectly paraphrasing 45% of slangs in Nigerian tweets and 24% for UK English, while ChatGPT and Gemini have fewer incorrect paraphrases. These results suggest that ChatGPT and Gemini handle slang more effectively due to their extensive training data and advanced architecture. In contrast, LLaMA3 struggles significantly with Nigerian slang, highlighting its potential limitations in understanding slang from specific cultural contexts. Given the lack of access for the research community to fine-tune commercial models like ChatGPT and Gemini, this reinforces the need for openly accessible models, such as LLaMA3, with improved slang knowledge.

5 Conclusion

In summary, our research evaluates the efficacy of LLMs in modelling slang, particularly in the context of climate-related tweets, though we speculate that our findings transfer to other topics. We observed significant emotional shifts when integrating slang paraphrases, in UK tweets and especially in Nigerian tweets. The shifts vary across LLaMA3, Gemini, and ChatGPT used for slang paraphrasing. Furthermore, factors like extensive training data and commercial nature likely contribute to ChatGPT's and Gemini's observed superiority on the task in comparison with LLaMA3. Our study highlights potential biases in LLMs towards non-Anglocentric regions and emphasizes the need for culturally-sensitive LLM development.

In future work, we plan to explore additional LLMs to facilitate a more comprehensive comparison of their performance in detecting and interpreting emotions in climate-related discourse. This will include evaluating newer models and their ability to understand regional slang and emotional nuances, as well as assessing their effectiveness across diverse cultural contexts. Additionally, we aim to enhance our dataset by incorporating real-time social media feeds to capture evolving slang and emotional expressions related to climate change. This expanded approach will provide deeper insights into how different LLMs process and represent emotional content in climate-related discussions.

6 Limitations

While our study underscores the importance of developing refined approaches to LLM development in diverse linguistic and cultural contexts, the reliance on a single model to zero-shot label emotions may limit the generalizability of the findings. Also, our research is constrained to specific demographic regions: Nigeria and the UK. To overcome these limitations, future studies should strive to incorporate multiple cultures and regions. Employing diverse methodologies will ensure a more comprehensive and nuanced analysis of emotional dynamics in discourse across global contexts.

7 Ethics Statement

The study followed the ACL Ethics Policy to ensure ethical and responsible conduct throughout the research process. We limited data gathering to publicly accessible tweets and anonymised the data to protect individuals privacy. Additionally, we avoid reinforcing biases or stereotypes and respectfully conduct the research in accordance with cultural norms and beliefs. The work makes use of suitable computational and statistical techniques, and we openly communicated our results to the larger scientific community. We are dedicated to maintaining moral standards in our studies.

8 Acknowledgements

The authors express gratitude to the Centre of Excellence for Data Science, Artificial Intelligence and Modelling (DAIM) and the Big Data Analytics (BDA) research 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

[1] Adedoja A Adedamola, Abiodun Modupe, and Olumuyiwa J Dehinbo. 2015. Development and Evaluation of a System for Normalizing Internet Slangs in Social Media Texts. In *Proceedings of*

the World Congress on Engineering and Computer Science 2015 Vol 1, San Francisco, USA. International Association of Engineers.

[2] AI@Meta. 2024. Llama 3 Model Card. Originaldate: 2024- 03- 15T17:57:00Z.

[3] 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.

[4] Yang Trista Cao, Anna Sotnikova, Jieyu Zhao, Linda X. Zou, Rachel Rudinger, and Hal Daumé III. 2024. Multilingual large language models leak human stereotypes across language boundaries. *arXiv preprint*. ArXiv:2312.07141 [cs].

[5] Ernie Chang, Jesujoba O. Alabi, David Ifeoluwa Adelani, and Vera Demberg. 2022. Few- Shot Pidgin Text Adaptation via Contrastive Fine-Tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4286–4291, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

[6] Yung- Sung Chuang, Mingye Gao, Hongyin Luo, James Glass, Hung- yi Lee, Yun- Nung Chen, and Shang- Wen Li. 2021. Mitigating Biases in Toxic Language Detection through Invariant Rationalization. *arXiv preprint*.

[7] 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.

[8] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. *arXiv preprint*. ArXiv:2104.08758 [cs].

[9] Constanza Fierro and Anders Søgaard. 2022. Factual Consistency of Multilingual Pretrained Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. Association for Computational Linguistics.

[10] Styliani Geronikolou, George Drosatos, and George Chrousos. 2021. Emotional Analysis of Twitter Posts During the First Phase of the COVID- 19 Pandemic in Greece: Infoveillance Study. *JMIR Formative Research*, 5(9):e27741.

[11] GoogleAI. 2024. Gemini Advanced - get access to Google's most capable AI model.

[12] Jochen Hartmann. 2022. Emotion English DistilRoBERTa- base.

[13] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and Strategies in Cross- Cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

[14] Bill Yuchen Lin, Frank F. Xu, Kenny Zhu, and Seungwon Hwang. 2018. Mining Cross- Cultural Differences and Similarities in Social Media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719, Melbourne, Australia. Association for Computational Linguistics.

[15] Maria L. Loureiro and Maria Alló. 2020. Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the U.K. and Spain. *Energy Policy*, 143(C). Publisher: Elsevier.

[16] Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The Biases of Pre- Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection. *IEEE Transactions on Affective Computing*, 14(3):1743–1753.

[17] Shamsuddeen Muhammad, Idris Abdulmumin, Abinew Ayele, Nedjma Ousidhoum, David Adelani, Seid Yimam, Ibrahim Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Alipio Jorge, Pavel Brazdil, Feleminio Ali, Davis David, Salomey Osei, Bello Shehu- Bello, Falalu Lawan, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Messelle, Hailu Balcha, Sisay Chala, Hagos Gebremichael, Bernard Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

[18] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting- Hao Huang, and Shomir Wilson. 2023. Nationality Bias in Text Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics.

[19] OpenAI. 2024. Introducing ChatGPT.

[20] Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. Slang Detection and Identification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 881–889, Hong Kong, China. Association for Computational Linguistics.

[21] Annelie Schoene, Piyusha Sanagavarapu, and Andrew Smith. 2020. Bidirectional dilated lstm with attention for fine- grained emotion classification in tweets. In *Affcon@ AAAI*, 2614, 100- 117.

[22] Yohei Seki and Yihong Liu. 2022. Multi- task Learning Model for Detecting Internet Slang Words with Two- Layer Annotation. In *2022 International Conference on Asian Language Processing (IALP)*, pages 212–218.

[23] 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.

[24] Laman R. Sultan. 2023. An Enhanced Emotion Classification Scheme for Twits Based on Deep Learning Approach. *Revue d'Intelligence Artificielle*, 37(5):1203–1211.

[25] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai- Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

[26] 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*.

[27] Zhewei Sun, Richard Zemel, and Yang Xu. 2021. A Computational Framework for Slang Generation. *Transactions of the Association for Computational Linguistics*, 9:462–478.

[28] Wei Wang, Ivan Hernandez, Daniel Newman, Jibo He, and Jiang Bian. 2016. Twitter Analysis: Studying US Weekly Trends in Work Stress and Emotion. *Applied Psychology*, 65:355–378.

[29] 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.

[30] Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji. 2023. Multilingual Large Language Models Are Not (Yet) Code- Switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Table 3: Emotion Distribution in Slang and Paraphrases of Climate Tweets from the UK and Nigeria. We used McNemar's test to determine if emotion categorisations changed significantly after paraphrasing. We applied a Bonferroni correction to account for multiple comparisons across categories. Bonferroni-adjusted significances are reported as: * : $p < 0.05/7 \simeq 0.0071$ ** $p < 0.01/7 \simeq 0.0014$ and ** $p < 0.001/7 \simeq 0.00014$

UK Climate Tweets Emotion Distribution					Nigeria Climate Tweets Emotion Distribution				
Sadness	Fear Anger	Neutral Surprise	Sadness Disgust	Anger	Surprise	Joy	Disgust	Fear	Neutral
Baseline Experiment (DistilRoberta)									
Tweets (slang)	550	436	314	360	304	264	251	68	6
28	8	13	0	5	1	8	5		
Tweets (paraphrased)									
External Knowledge Integration Experiments									
Llama3	607^{***}	104^{***}	449^{***}	247^{***}	185^{***}	656^{***}	5^{***}	14	9^{***}
1	2	9^{***}	6	1	5	0	1	8	8^{***}
ChatGPT-4	687^{***}	110^{***}	437^{***}	189^{***}	245^{*}	582^{***}	3^{***}	16	6^{***}
1	6	4^{***}	1	6	0^{*}	5	0	4	1
1	6	0^{***}							
Gemini	648^{***}	108^{***}	456^{***}	225^{***}	245^{*}	567^{***}	4^{***}	14	6^{***}
8^{*}	1	1	9^{*}	8	4^{***}	5	0	1	8
4^{***}									

Table 4: Examples of emotion changes given slang and paraphrases across LLMs.

Climate tweets with slangs	Slang	Origin	Gold label	Baseline Emotion	ChatGPT		Gemini		Llama	
					Paraphrase	Emotion	Paraphrase	Emotion	Paraphrase	Emotion
"can you imagine your guy fall hand on the climate change question o"	"fall hand"	Nigeria	"disappoint"	Fear	"to disappoint"	Sadness	"to disappoint"	Sadness	"Fall from grace"	Joy
"not in the united kingdom it isn't global warming my arse its more like autumn or fall as you guys call it"	"arse"	UK	"buttocks"	Sadness	"buttocks"	Anger	"buttocks"	Anger	"derogatory"	Neutral
"rain in november na wa this climate change is real"	"na wa"	Nigeria	"damn"	Fear	"wow"	Surprise	"oh no!"	Surprise	"unfortunate"	Fear
"invenus what sort of bloke is that in the present climate works on kabul managing a hotel"	"bloke"	UK	"a man"	Surprise	"a man"	Fear	"a man"	Fear	"a man"	Fear