BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages

Shamsuddeen Hassan Muhammad[1,2*], **Nedjma Ousidhoum**[3*],
Idris Abdulmumin[4], **Jan Philip Wahle**[5], **Terry Ruas**[5], **Meriem Beloucif**[6], **Christine de Kock**[7], **Nirmal Surange**[8],
Daniela Teodorescu[9], **Ibrahim Said Ahmad**[10], **David Ifeoluwa Adelani**[11,12,13], **Alham Fikri Aji**[14],
Felermino D. M. A. Ali[15], **Ilseyar Alimova**, **Vladimir Araujo**[16], **Nikolay Babakov**[17], **Naomi Baes**[7],
Ana-Maria Bucur[18,19], **Andiswa Bukula**[20], **Guanqun Cao**[21], **Rodrigo Tufino Cardenas**[22], **Rendi Chevi**[14],
Chiamaka Ijeoma Chukwuneke[23], **Alexandra Ciobotaru**[18], **Daryna Dementieva**[24], **Murja Sani Gadanya**[2],
Robert Geislinger[25], **Bela Gipp**[5], **Oumaima Hourrane**[26], **Oana Ignat**[27], **Falalu Ibrahim Lawan**[28],
Rooweither Mabuya[20], **Rahmad Mahendra**[29], **Vukosi Marivate**[4], **Andrew Piper**[12], **Alexander Panchenko**,[30],
Charles Henrique Porto Ferreira[31], **Vitaly Protasov**, **Samuel Rutunda**[32], **Manish Shrivastava**[8],
Aura Cristina Udrea[33], **Lilian Diana Awuor Wanzare**[34], **Sophie Wu**[12], **Florian Valentin Wunderlich**[5],
Hanif Muhammad Zhafran[35], **Tianhui Zhang**[36], **Yi Zhou**[3],
Saif M. Mohammad[37]

[1]Imperial College London, [2]Bayero University Kano, [3]Cardiff University, [4]DSFI, University of Pretoria, [5]University of Göttingen
[6]Uppsala University, [7]University of Melbourne, [8]IIIT Hyderabad, [9]University of Alberta, [10]Northeastern University,
[11]MILA, [12]McGill University, [13]Canada CIFAR AI Chair,[14]MBZUAI,[15]LIACC, FEUP, University of Porto, [16]Sailplane AI,
[17]University of Santiago de Compostela, [18]University of Bucharest, [19]Universitat Politècnica de València, [20]SADiLaR,
[21]University of York, [22]Universidad Politécnica Salesiana,[23]Lancaster University, [24]Technical University of Munich,[25]Hamburg University,
[26]Al Akhawayn University, [27]Santa Clara University, [28]Kaduna State University, [29]Universitas Indonesia, ,[30]Skoltech, [31]Centro Universitário FEI,
[32]Digital Umuganda, [33]National University of Science and Technology Politehnica Bucharest, [34]Maseno University, [35]Institut Teknologi Bandung,
[36]University of Liverpool,[37]National Research Council Canada

Contact: s.muhammad@imperial.ac.uk, OusidhoumN@cardiff.ac.uk

Abstract

People worldwide use language in subtle and complex ways to express emotions. While emotion recognition – an umbrella term for several NLP tasks – significantly impacts different applications in NLP and other fields, most work in the area is focused on high-resource languages. Therefore, this has led to major disparities in research and proposed solutions, especially for low-resource languages that suffer from the lack of high-quality datasets. In this paper, we present BRIGHTER – a collection of multilabeled emotion-annotated datasets in 28 different languages. BRIGHTER covers predominantly low-resource languages from Africa, Asia, Eastern Europe, and Latin America, with instances from various domains annotated by fluent speakers. We describe the data collection and annotation processes and the challenges of building these datasets. Then, we report different experimental results for monolingual and crosslingual multi-label emotion identification, as well as intensity-level emotion recognition. We investigate results with and without using LLMs and analyse the large variability in performance across languages and text domains. We show that BRIGHTER datasets are a step towards bridging the gap in text-based emotion recognition and discuss their impact and utility.

1 Introduction

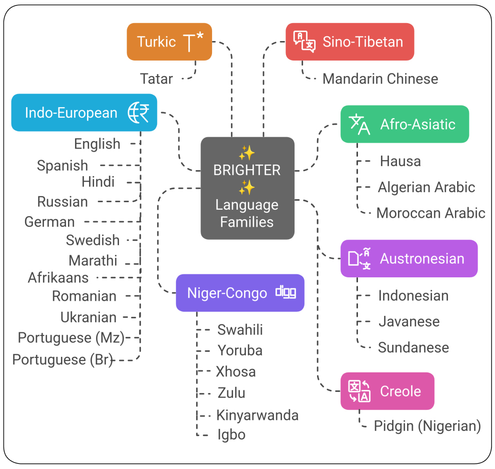


Figure 1: Languages included in BRIGHTER and their language families.

While emotions are expressed and managed daily, they are complex, nuanced, and sometimes hard to articulate and interpret. That is, people use language in subtle and complex ways to express emotions across languages and cultures (Wiebe et al., 2005; Mohammad and Kiritchenko, 2018; Mohammad et al., 2018a) and perceive them subjectively, even within the same culture or social

[*]Equal contribution

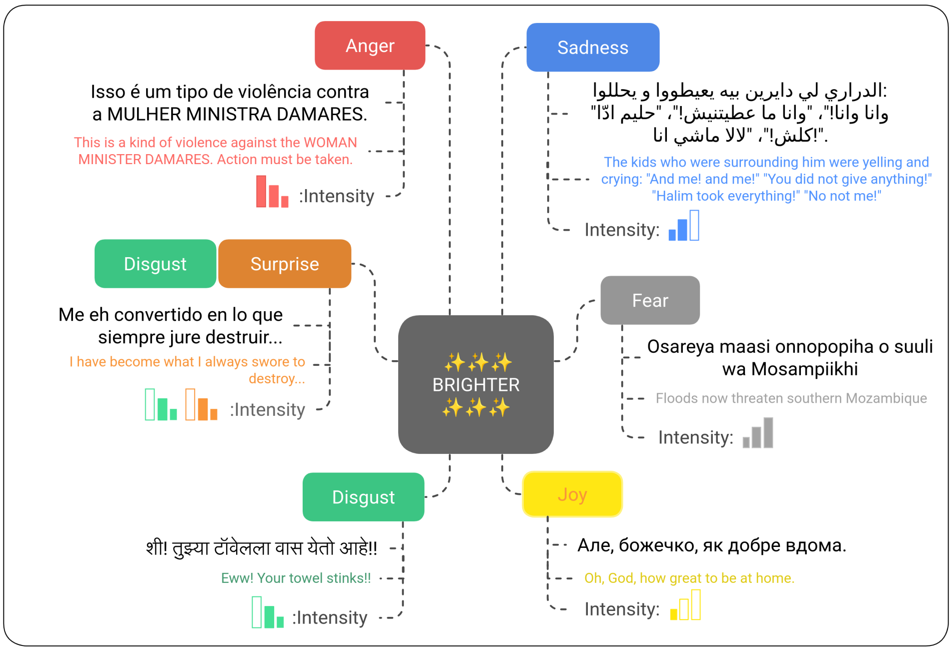

Anger
Isso é um tipo de violência contra a MULHER MINISTRA DAMARES.
This is a kind of violence against the WOMAN MINISTER DAMARES. Action must be taken.
:Intensity
Sadness
الدراري لي دايرين بيه يعيطووا و يحللوا:
"وانا وانا!"، "وانا ما عطيتنيش!"، "حليم ادّا
كلش!"، "لالا ماشي انا".
The kids who were surrounding him were yelling and crying: "And me! and me!" "You did not give anything!" "Halim took everything!" "No not me!"
Intensity:
Disgust
Surprise
Me eh convertido en lo que siempre jure destruir...
I have become what I always swore to destroy...
:Intensity
✨✨✨ BRIGHTER ✨✨✨
Fear
Osareya maasi onnopopiha o suuli wa Mosampiikhi
Floods now threaten southern Mozambique
Intensity:
Disgust
शी! तुझ्या टॉवेलला वास येतो आहे!!
Eww! Your towel stinks!!
:Intensity
Joy
Але, божечко, як добре вдома.
Oh, God, how great to be at home.
Intensity:

Figure 2: **Examples from the BRIGHTER dataset collection** in 6 different languages with their translations and intensity levels. Note that the instances can have one or more labels (e.g., disgust and surprise as shown in the figure.)

group. Emotion recognition is at the core of several NLP applications in healthcare, dialogue systems, computational social science, digital humanities, narrative analysis, and several others (Mohammad et al., 2018b; Saffar et al., 2023). It is an umbrella term for multiple NLP tasks, such as detecting the emotions of the speaker, identifying what emotion a piece of text is conveying, and detecting the emotions evoked in a reader (Mohammad, 2022). In this paper, we use *emotion recognition* to refer to *perceived* emotions, i.e., what emotion most people think the speaker might have felt given a sentence or a short text snippet uttered by the speaker.

Most work on emotion recognition has focused on high-resource languages such as English, Spanish, German, and Arabic (Strapparava and Mihalcea, 2007; Seyeditabari et al., 2018; Chatterjee et al., 2019; Kumar et al., 2022). This is partly due to the unavailability of datasets in under-served languages, which has led to a major research gap in the area, which is particularly noticeable in low-resource languages. That is, despite the linguistic diversity present in different parts of the world, such as Africa and Asia, which are home to more than 4,000 languages[1], few emotion recognition resources are available in these languages. To bridge this gap, we introduce BRIGHTER – a collection of manually annotated emotion datasets for 28 languages containing nearly 100,000 instances from diverse data sources: speeches, social media, news, literature, and reviews. The languages belong to 7 language families (see Figure 1) and are predominantly low-resource, mainly spoken in **Africa**, **Asia**, **Eastern Europe**, **Latin America**, along with mid- to high-resource languages such as English. Each instance in BRIGHTER is curated and annotated by fluent speakers based on six emotion classes: *joy, sadness, anger, fear, surprise, disgust, and none*. The instances are multi-labeled and include 4 levels of intensity that vary from 0 to 3 (examples in Figure 2). We describe the collection, annotation, and quality control steps used to construct BRIGHTER. We then test various baseline experiments and observe that LLMs still struggle with recognising perceived emotions in text. We further report on the observed discrepancies across languages such as the fact that, for low-resource languages, LLMs perform significantly better when prompted in English. We make our datasets public[2], which presents an important step towards work on emotion recognition and related tasks as we involve local communities in the collection and annotation.

[1] https://www.ethnologue.com/insights/how-many-languages

[2] https://github.com/emotion-analysis-project/SemEval2025-Task11

Our insights into language-specific characteristics of emotions in text, nuances, and challenges may enable the creation of more inclusive digital tools.

2 The BRIGHTER Dataset Collection

2.1 Data Collection

As our BRIGHTER collection includes 28 different datasets, curated and annotated by fluent speakers, we use different data sources, collection, and annotation strategies depending on 1) the availability of the textual data potentially rich in emotions and 2) access to annotators. We detail the various choices made when selecting and balancing data sources, annotating the instances, and controlling for data quality in the following section.

2.1.1 Data Sources

Choosing suitable data sources is challenging when resources are lacking. Therefore, we typically combine data sources as shown in Table 1. We present the main textual sources we used to build BRIGHTER in the following.

Social media posts We use social media data collected from various platforms, including Reddit (e.g., eng, deu), YouTube (e.g., esp, ind, jav, sun), Twitter (e.g., hau, ukr), and Weibo (e.g., chn). For some languages, we re-annotate existing sentiment datasets for emotions (e.g., the sentiment analysis benchmark AfriSenti (Muhammad et al., 2023a) for ary, hau, kin; the Twitter dataset by Bobrovnyk (2019) for ukr).

Personal narratives, talks, speeches Anonymised sentences from personal diary posts are ideal for extracting sentences where the speaker is centering their own emotions as opposed to the emotions of someone else. Hence, we use these in eng, deu, and ptbr, mainly from subreddits such as, e.g., IAmI.

Similarly, the afr dataset includes sentences from speeches and talks which constitute a good source for potentially emotive text.

Literary texts We translated the novel *"La Grande Maison"* (The Big House) by Mohammed Dib [3] from French to Algerian Arabic and post-processed the translation to generate sentences to be annotated by native speakers. Note that the translator is bilingual and a native Algerian Arabic speaker. Such a source is typically rich in emotions as it includes interactions between various characters. Further, Algerian Arabic is mainly spoken due to the Arabic diglossia, which makes this resource valuable since it highly differs from social media datasets in arq.

News data Although we prefer emotionally rich social media data from different platforms, such data is not always available. Therefore, to collect a larger number of instances, we annotate news data and headlines in some African languages (e.g., yor, hau, and vmw).

Human-written and machine generated data We create a dataset from scratch for Hindi (hin) and Marathi (mar). We ask annotators to generate emotive sentences on a given topic (e.g., family). In addition, we automatically translate a small section of the Hindi dataset to Marathi, and native speakers manually fix the translation errors. Finally, we augment both datasets with a few hundred quality-approved instances generated by ChatGPT.

2.1.2 Pre-processing and Quality Control

Prior to annotation, we preprocess the data by removing duplicates, invisible characters, garbled encoding, and incorrectly rendered emoticons. We anonymise all texts and exclude content with excessive expletives or dehumanising language.

2.2 Annotating BRIGHTER

As a text snippet can elicit multiple emotions simultaneously, we ask the annotators to select all the emotions that apply to a given text rather than choosing a single dominant emotion class. The set of labels includes six categories of perceived emotions: *anger, sadness, fear, disgust, joy, surprise*, and *neutral (if no emotion is present)*. The annotators further rate the selected emotion(s) on a four-point intensity scale: 0 (no emotion), 1 (low intensity), 2 (moderate intensity level), and 3 (high intensity). We provide the definitions of the categories and annotation guide in Appendix B.

We use Amazon Mechanical Turk to annotate the English dataset, and Toloka to label the Russian, Ukrainian, and Tatar instances. However, as traditional crowdsourcing platforms do not have a large pool of annotators who speak various low-resource languages, we directly recruit fluent speakers to annotate the data and use the academic version of LabelStudio (Tkachenko et al., 2020-2025) and Potato (Pei et al., 2022) to set up our annotation platform.

[3] https://en.wikipedia.org/wiki/La_Grande_Maison

Language	Data source(s)	#Annotators (total)	#Ann. / sample	Train	Dev	Test	Total
Afrikaans (**afr**)	Speeches	3	3	2,107	98	1,065	3,270
Algerian Arabic (**arq**)	Literature	10	4 to 9	901	100	902	1,903
Moroccan Arabic (**ary**)	News, social media	3	3	1,608	267	812	2,687
Chinese (**chn**)	Social media	7	5	2,642	200	2,642	5,484
German (**deu**)	Social media	10	7	2,603	200	2,604	5,407
English (**eng**)	Social media	122	5 to 30	2,768	116	2,767	5,651
Latin American Spanish (**esp**)	Social media	12	5	1,996	184	1,695	3,875
Hausa (**hau**)	News, social media	5	5	2,145	356	1,080	3,581
Hindi (**hin**)	Created	5	4 to 5	2,556	100	1,010	3,666
Igbo (**ibo**)	News, social media	3	3	2,880	479	1,444	4,803
Indonesian (**ind**)	Social media	16	3	–	156	851	1,007
Javanese (**jav**)	Social media	13	3	–	151	837	988
Kinyarwanda (**kin**)	News, social media	3	3	2,451	407	1,231	4,089
Marathi (**mar**)	Created	4	4	2,415	100	1,000	3,515
Nigerian-Pidgin (**pcm**)	News, social media	3	3	3,728	620	1,870	6,218
Portuguese (Brazilian; **ptbr**)	Social media	5	5	2,226	200	2,226	4,652
Portuguese (Mozambican; **ptmz**)	News, social media	3	3	1,546	257	776	2,579
Romanian (**ron**)	Social media	8	3 to 8	1,241	123	1,119	2,483
Russian (**rus**)	Social media	10	3 to 10	2,679	199	1,000	3,878
Sundanese (**sun**)	Social media	16	3	924	199	926	2,049
Swahili (**swa**)	News, social media	3	3	3,307	551	1,656	5,514
Swedish (**swe**)	Social media	3	3	1,187	200	1,188	2,575
Tatar (**tat**)	Social media	3	2	1,000	200	1,000	2,200
Ukrainian (**ukr**)	Social media	106	5	2,466	249	2,234	4,949
Emakhuwa **vmw**	News, social media	3	3	1,551	258	777	2,586
isiXhosa (**xho**)	News, social media	3	3	–	682	1,594	2,276
Yoruba (**yor**)	News	3	3	2,992	497	1,500	4,989
isiZulu (**zul**)	News, social media	3	3	–	875	2,047	2,922

Table 1: BRIGHTER data sources, annotator counts and data splits, sorted alphabetically by language code. Datasets with no training splits (-) were only used for testing (see Section 3).

2.3 Annotators' Reliability

While both inter-annotator agreement (IAA) and reliability scores measure the quality of annotations, they address different aspects. That is, IAA evaluates how much the annotators agree with each other, whereas reliability scores focus on the consistency of the aggregated labels across different trials (repeated annotations; Kiritchenko and Mohammad, 2016). Hence, when the final aggregated labels are obtained from a larger number of annotations, reliability scores tend to increase. In contrast, IAA scores do not depend on the number of annotations per instance. We report the reliability of the annotation using Split-Half Class Match Percentage (SHCMP; Mohammad, 2024). SHCMP extends the concept of Split-Half Reliability (SHR), traditionally used for continuous scores (Kiritchenko and Mohammad, 2016), to discrete categories like ours (i.e., intensity scores per emotion). SHCMP measures the extent to which n bins (i.e., subsets corresponding to halves when $n = 2$) of the annotations classify items in the same way by splitting a dataset with individual labels into n random bins, and computing how many times each item in each bin is assigned the same class or category. This calculation is repeated 1,000 times and the average corresponds to the final SHCMP score. That is, a higher SHCMP indicates that repeated annotations would produce similar class labels (i.e., higher reliability). Further explanations can be found in Appendix B. Figure 3 shows a heatmap presenting the SHCMP values for the BRIGHTER datasets. Overall, the SHCMP scores are high ($> 60\%$ for $n = 2$), which indicates that our annotations are reliable.

2.4 Determining the Final Labels

We expected a level of disagreement as emotions are complex, subtle, and perceived differently even from people within the same culture. In addition, text-based communication is limited as it lacks cues such as tone, relevant context, and information about the speaker. Our approach for aggregating the per-annotator emotion and intensity labels is detailed below. We also publicly share the individual (non-aggregated) annotations, recognising that annotator disagreement can provide useful signals in itself (Plank, 2022).

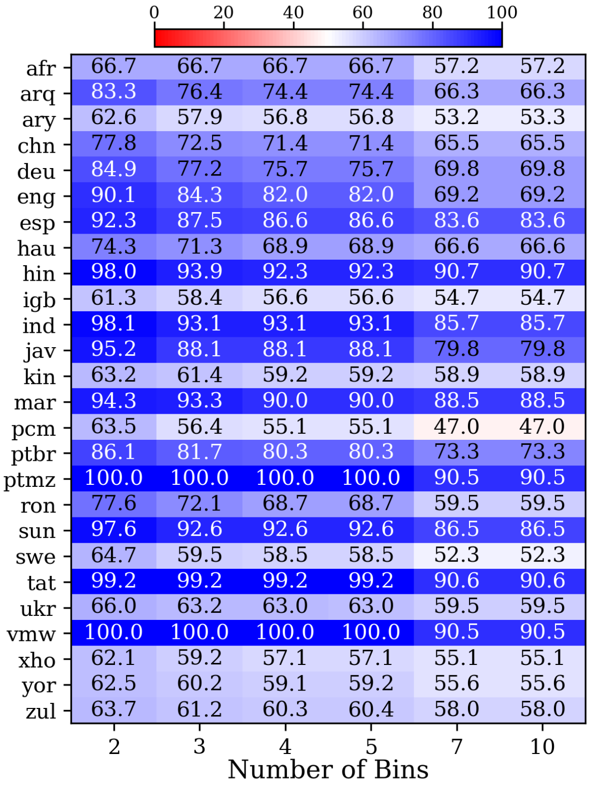


Figure 3: SHCMP (%) values for the BRIGHTER datasets across varying numbers of bins (2 to 10). Higher values indicate better reliability scores. Note that ptmz and vwm have the same score as vwm instances were translated from ptmz and the translation was verified.

Aggregating the Emotion Labels The final emotion labels are determined based on the emotions and associated intensity values selected by the annotators. That is, the given emotion is considered present if:

1. At least two annotators select a label with an intensity value of 1, 2, or 3 (low, medium, or high, respectively).
2. The average score exceeds a predefined threshold T. We set T to 0.5.

Aggregating the Intensity Labels Once the labels for perceived emotions are assigned, we determine the final intensity score for each instance by averaging the selected intensity scores and choosing the ceiling. We only assign intensity scores for datasets where most instances are annotated by ≥ 5 annotator to ensure robustness.

2.5 Final Data Statistics

Figure 4 shows the distribution of the annotated emotions in the BRIGHTER datasets. The neutral class contains instances that do not belong to any of the six predefined categories (i.e., anger, disgust, fear, sadness, joy, and surprise). Although most languages include all six categories, the English dataset does not include disgust, and the Afrikaans one does not include surprise due to an insufficient class representation. Furthermore, class distributions show substantial variation as we chose various data sources as shown in Table 1.

3 Experiments

3.1 Setup

We report the data split sizes in Table 1. The test sets are large, with about 1,000 instances and up to almost 3,000. Datasets with no training data are not used for training. For our baseline experiments, we test multi-label emotion classification and emotion intensity prediction using Multilingual Language Models (MLMs) and Large Language Models (LLMs) for the following.

Multi-label Emotion Classification in Few-shot Settings We report the emotion classification performance using five LLMs–QWEN2.5-72B (Yang et al., 2024), DOLLY-V2-12B (Conover et al., 2023), LLAMA-3.3-70B (Touvron et al., 2023), MIXTRAL-8X7B (Jiang et al., 2024), and DEEPSEEK-R1-70B (DeepSeek-AI et al., 2025). We prompt the LLMs to perform Chain-of-Thought (CoT) and predict the presence of each emotion from a predefined set, set the number of few-shot examples to 8, and consider the first answer generated by the LLMs (i.e., top–1). We report the macro F1-score results on 28 languages. In Appendix B, we also report monolingual classification results for all 24 languages with training datasets in Table 5.

Multi-label Emotion Classification in Crosslingual Settings We report the macro F-score results for systems trained without using any data in the 28 target languages when testing on each. Hence, we train MLMs on all languages in one family (see Figure 1) except for one held-out target language, which we test on and report the results for each test set. For families with only one language, we train on Slavic languages (rus and ukr) and test on tat; two Niger-Congo languages (swa and yor) and test on pcm; and on rus and test on chn.

Emotion Intensity Prediction We report the Pearson correlation scores for systems trained on

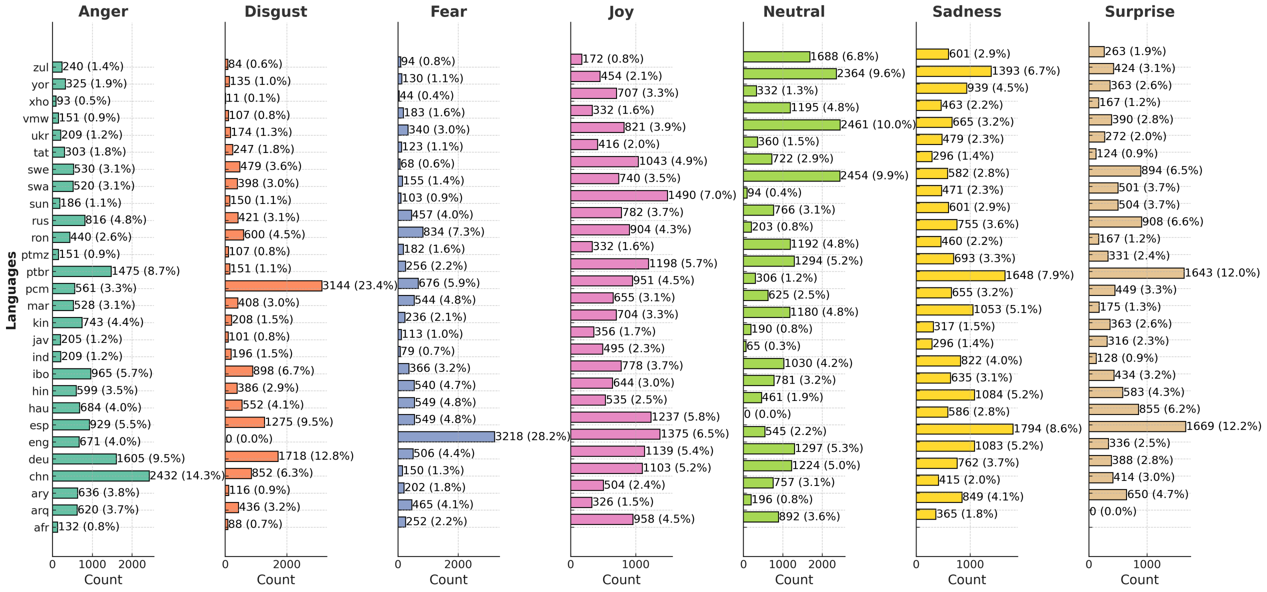


Figure 4: Emotion label distribution across BRIGHTER datasets. Each bar represents the number of labeled instances per emotion (i.e., anger, disgust, fear, joy, sadness, surprise, and neutral) and its percentage.

the intensity-labeled training sets in 10 languages.

3.2 Experimental Results

Table 2 reports the results of few-shot and crosslingual experiments for multi-label emotion classification and Table 3 reports those for emotion intensity classification. Our results corroborate how challenging emotion classification is for LLMs, even for high-resource languages such as eng and deu. The performance is worse for low-resource languages, for which Dolly-v2-12B performs the worst, and Qwen2.5-72B performs the best on average.

We observe the largest performance for yor with a maximum of 27.44. hin, mar, and tat have the best performance among all languages, which is unsurprising since the tat dataset is single-labeled, and close to 70% and 80% of the test data for mar and hin respectively are single-labeled.

Multi-label Emotion Recognition Results The crosslingual experiments show that the model performance depends on both the languages used for the transfer learning and those used for pretraining the LLM. For instance, in some languages, training on other languages from the same family boosts the performance and outperforms the few-shots settings (e.g., swe when RemBERT is fine-tuned on Germanic languages). However, all the Niger-Congo languages (vmw in particular) are those that benefited the least from the crosslingual transfer across all models, with RemBERT performing the worst. This is largely due to their under-resourcedness even when combining data. Notably, XLM-R performs exceptionally well in languages such as deu, chn, hin, ptbr, but struggles significantly in others (e.g., swe, ptmz). In contrast, mDeBERTa's results are the most stable across most languages with low scores for ibo, vmw, and yor which are not part of the CC-100 corpus (Conneau et al., 2020) used for training mDeBERTa. One would also argue that mDeBERTa was also not trained on arq but the Modern Standard Arabic (MSA) data used for training the model helped boost the performance.

Emotion Intensity For the intensity detection, which is more challenging, Dolly-v2-12B's results are worse whereas DeepSeek-R1-70B shows promising results by outperforming other models in most languages. Interestingly, MLMs achieve better results, notably RemBERT on high-resource languages (deu, eng, esp, rus) with chn being the only exception. On the other hand, for mainly vernacular (i.e., spoken) low-resource (e.g., arq) LLMs show some striking improvements (>36 with DeepSeek-R1-70B).

4 Analysis

The results in Figure 5a suggest that LLM performance is highly dependent on the prompt wording when asking for the presence of emotion on the English test set using different paraphrases of the

	Few-Shot Multi-Label Classification					Crosslingual Multi-Label Classification				
Lang.	Qwen2.5-72B	Dolly-v2-12B	Llama-3.3-70B	Mixtral-8x7B	DeepSeek-R1-70B	LaBSE	RemBERT	XLM-R	mBERT	mDeBERTa
afr	60.18	23.58	**61.28**	53.69	43.66	35.12	35.04	**41.66**	16.95	33.25
arq	37.78	38.59	**55.75**	45.29	50.87	**35.93**	33.78	35.87	31.38	35.92
ary	**52.76**	24.27	44.96	35.07	47.21	**42.83**	35.46	33.88	24.83	36.28
chn	**55.23**	27.52	53.36	44.91	53.45	45.28	24.56	**53.84**	21.61	42.41
deu	**59.17**	26.86	56.99	51.20	54.26	42.45	46.84	**47.26**	28.60	42.61
eng	55.72	42.60	**65.58**	58.12	56.99	36.71	37.54	**37.60**	18.80	35.30
esp	72.33	36.41	61.27	65.72	**73.29**	54.56	**57.37**	44.52	30.09	37.09
hau	43.79	29.43	50.91	40.40	**51.91**	**38.46**	31.98	16.69	15.59	32.80
hin	**79.73**	27.59	60.59	62.19	76.91	69.78	13.75	**69.96**	36.94	57.74
ibo	**37.40**	24.31	33.18	31.90	32.85	**18.13**	7.49	10.42	9.94	9.52
ind	**57.29**	36.61	39.20	54.37	49.51	**47.50**	37.64	25.39	26.87	35.68
jav	**50.47**	36.18	41.88	48.37	43.05	46.24	**46.38**	20.39	26.16	35.34
kin	31.96	19.73	**34.36**	26.35	32.52	**30.35**	18.38	13.12	20.90	17.30
mar	74.58	25.69	67.40	50.36	**76.68**	74.65	**77.24**	76.21	42.32	54.05
pcm	38.66	34.41	**48.67**	45.61	45.00	**33.29**	1.01	21.08	22.55	25.39
ptbr	**51.60**	25.90	45.03	41.64	51.49	41.51	41.84	**43.09**	23.86	34.42
ptmz	**40.44**	16.70	34.06	36.52	39.58	**31.44**	29.67	7.30	13.54	24.46
ron	68.18	43.58	**71.28**	68.51	65.02	69.79	**76.23**	65.21	61.50	60.60
rus	73.08	29.72	62.61	61.72	**76.97**	61.32	**70.43**	21.14	37.15	29.70
sun	42.67	32.20	**46.33**	42.10	44.61	**34.79**	19.43	25.92	25.29	27.31
swa	27.36	17.63	29.47	26.51	**33.27**	21.66	**18.99**	16.94	18.61	14.94
swe	48.89	21.79	**50.26**	48.61	44.60	44.24	**51.18**	10.08	28.86	43.28
tat	51.58	25.12	49.84	39.44	**53.86**	**60.66**	44.54	39.58	35.81	47.72
ukr	**54.76**	17.16	42.34	40.15	51.19	44.37	**49.56**	34.06	25.69	35.12
vmw	**20.41**	16.03	18.96	19.00	19.09	9.65	5.22	**12.66**	12.11	11.74
xho	29.56	24.12	30.79	22.92	29.08	31.39	12.73	11.48	17.08	22.86
yor	24.99	16.00	23.70	19.67	**27.44**	**11.64**	5.33	6.64	9.62	10.03
zul	22.03	14.72	21.48	20.38	20.38	18.16	15.26	10.92	13.04	13.87
AVG	**49.71**	26.88	47.12	43.56	49.21	**40.50**	33.63	30.61	24.16	32.38

Table 2: Average F1-Macro for multi-label emotion classification. In the few-shot setting, we predict the emotion class on test set in 28 languages. In the crosslingual setting, we train on all languages within a language family except the target language, and evaluate on the test set of the target language. The best performance scores in monolingual and zero-shot settings are highlighted in **blue** and **orange**, respectively.

same text. Further, Figure 5b shows that, when testing the effect of n-shot settings on the English test set, we observe a significant improvement in performance with more shots, with Mixtral-8x7B and Llama-3.3-70B outperforming other models. However, the scores tend to reach a plateau at 4 shots for all LLMs except for Qwen2.5-72B, which suggests that 4 to 8 shots may be sufficient to obtain stable results. In addition, when testing how likely we can get the correct answer when prompting LLMs to generate tokens based on a top-k selection, the results shown in Figure 5c suggest that increasing the value of k results consistently in better performance, particularly when using DeepSeekR1-70B, which achieves an F-score > 90 when $k = 8$.

When comparing the performance of the models prompted in English vs. the target language, Figure 6 shows that LLMs tend to perform better when prompted in English except for arq for which Qwen2.5-72B performs better when prompted in MSA. Improvements when using English prompts are markedly noticeable in low-resource languages (e.g., hau, mar, vmw) where Dolly-v2-12B and LLamaa-3.3-70B struggle the most with target language prompts.

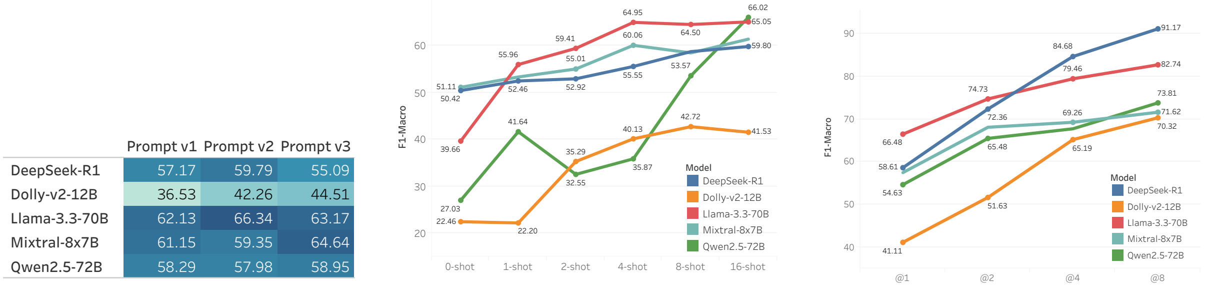
5 Related Work

Appraisal theories of emotion describe that emotions are due to our evaluation of an event based on personal experiences, resulting in various emotions evoked for different people (Arnold, 1960; Moors et al., 2013; Ellsworth, 2013; Frijda, 1986; Lazarus, 1991; Ortony et al., 2022; Roseman, 2013; Scherer, 2009). The theory of constructed emotions states that they are not hard-wired in the brain or universal, but are rather concepts constructed by the brain (Barrett, 2016, 2017).

Prior work in NLP has largely focused on *sentiment analysis* – detecting whether a text expresses positive, negative, or neutral valence (Mohammad, 2016; Muhammad et al., 2023b). Recent work focus has shifted to a broader form—detecting emo-

	Multilingual Language Models (MLMs)					Large Language Models (LLMs)				
Lang.	LaBSE	RemBERT	XLM-R	mBERT	mDeBERTa	Qwen2.5-72B	Dolly-v2-12B	Llama-3.3-70B	Mixtral-8x7B	DeepSeek-R1-70B
arq	1.42	**1.64**	0.89	1.10	0.47	29.54	3.80	36.29	31.05	**36.37**
chn	23.37	**40.53**	36.92	21.96	23.25	46.17	8.11	**51.86**	46.52	48.57
deu	28.93	**56.21**	38.30	17.35	18.14	43.30	7.43	53.46	47.60	**54.78**
eng	35.34	**64.15**	37.36	25.74	8.85	**55.99**	13.35	44.14	55.26	48.08
esp	56.89	**72.59**	55.72	27.94	29.18	51.11	10.49	51.64	55.54	**60.74**
hau	26.13	**27.03**	24.68	2.79	0.00	27.00	6.43	**39.16**	25.84	38.85
ptbr	20.62	**29.74**	18.24	8.36	1.32	38.20	9.02	40.90	39.17	**46.72**
ron	35.57	**55.66**	37.77	21.99	4.63	55.48	12.62	45.87	57.07	**57.69**
rus	68.43	**87.66**	68.96	37.63	5.03	58.25	13.96	57.56	56.01	**62.28**
ukr	13.75	**39.94**	36.16	4.32	3.51	37.74	6.04	36.99	38.74	**43.54**
AVG	30.54	**46.61**	35.25	16.35	9.97	43.03	8.74	45.78	43.97	**48.88**

Table 3: Pearson correlation scores for intensity classification using MLMs and LLMs. The best performance scores are highlighted in **blue** and **orange**, respectively.



(a) **Performance of different LLMs across three prompt paraphrases on the English test set.** Different prompts impact model performance.

(b) **Few-shot performance of LLMs on the English test set.** Performance improves with more shots.

(c) **Top-k performance of different LLMs on the English test set.** Higher k values increase the likelihood of retrieving the correct answer.

Figure 5: Ablation studies on the effect of prompt wording variation, few-shot examples, and top-k predictions conducted on the English test set.

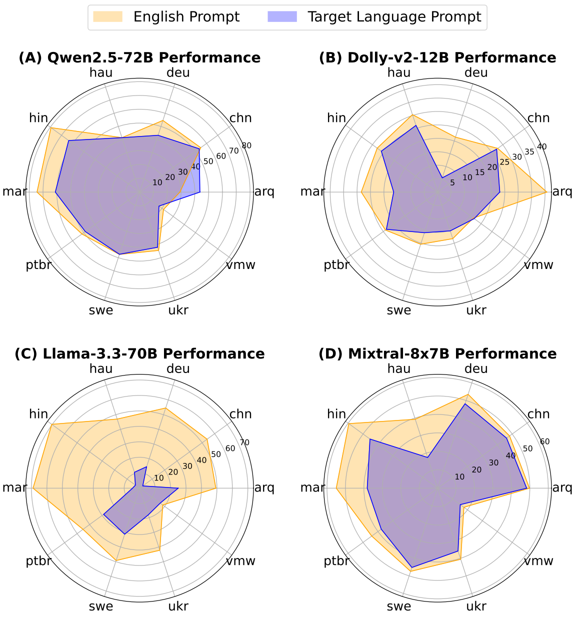


Figure 6: **Comparing models' performance across languages when prompted in English (orange) vs. when prompted in the target language (blue)**. LLMs perform better when prompted in English.

tions in text such as anger, fear, joy, sadness, etc. in text which is in line with discrete models of emotions (e.g., Paul Eckman's six emotions (Ekman, 1992) and Plutchik's Wheel of Emotions (Plutchik, 1980) for anger, disgust, fear, happiness, sadness, surprise, anticipation and trust).

Several initiatives have created emotion classification datasets for languages other than English (e.g., Italian (Bianchi et al., 2021), Romanian (Ciobotaru and Dinu, 2021), Indonesian (Saputri et al., 2018), and Bengali (Iqbal et al., 2022)). However, NLP work in the area is predominantly Western-centric, and while multilingual datasets like XED (Öhman et al., 2020) and XLM-EMO (Bianchi et al., 2022) exist, XLM-EMO's reliance on translated data for over ten languages may not fully capture cultural nuances in emotion expression. Emotions are culture-sensitive and highly contextualized, influenced by cultural values (Havaldar et al., 2023; Mohamed et al., 2024; Hershcovich et al., 2022). Further, although emotions can cooccur (Vishnubhotla et al., 2024), most datasets assume single-label classification. While GoEmotions (Demszky et al., 2020) addresses multi-label emotion classification, to our knowledge, no multilingual resources capture the overlapping emotions and intensity across languages. This work aims to push this boundary by presenting emotion-labeled data for 28 languages. Given the lack of unanimity surrounding language categorisation as low-resource, approximately 15 to 17 of these languages could be considered such.

6 Conclusion

We presented BRIGHTER, a collection of emotion recognition datasets in 28 languages spoken across various continents. The instances in BRIGHTER are multi-labeled, collected, and annotated by fluent speakers, with 10 datasets annotated for emotion intensity. When testing LLMs on our dataset collection, the results show that they still struggle with predicting perceived emotions and their intensity levels, especially for under-resourced languages. Further, our results show that LLM performance is highly dependent on the wording of the prompt, its language, and the number of shots in few-shot settings. We publicly release BRIGHTER, our annotation guidelines, and individual labels to the research community.

Limitations

Emotions are subjective, subtle, expressed, and perceived differently. We do not claim that BRIGHTER covers the true emotions of the speakers, is fully representative of the language use of the 28 languages, or covers all possible emotions. We discuss this extensively in the Ethics Section.

We are aware of the limited data sources in some low-resource languages. Therefore, our datasets cannot be used for tasks that require a large amount of data from a given language. However, they remain a good starting point for research in the area.

Ethical Considerations

Emotion perception and expression are subjective and nuanced as they are strongly related to a myriad of other aspects (e.g., cultural background, social group, personal experiences, social context). Thus, we can never truly identify how one is feeling based solely on shot text snippets with absolute certainty. We clearly state that our datasets focus on perceived emotions and determining what emotion most people think the speaker may have felt. Hence, we do not claim that we annotate the true emotion of the speaker, which cannot be definitively known from just a short text snippet. We acknowledge the importance of this distinction as perceived emotions can differ from actual emotions.

We acknowledge possible biases in our data since we rely on text-based communication, where data sources can include biases, and annotators might always come with their own internalized subtle ones. Further, although many of our datasets focus on low-resource languages, we do not claim that they fully represent these languages' usage, and while we controlled for inappropriate instances, we acknowledge that we might have missed some.

We explicitly ask for careful reflection on the ethical considerations before using our datasets. We forbid using our datasets for commercial purposes or by state actors to make high-risk applications unless explicitly approved by the dataset creators. Systems built on our systems may not be reliable at individual instance levels and are impacted by domain shifts. Therefore, they should not be used to make critical decisions for individuals, such as in health applications without expert supervision. See Mohammad (2022, 2023) for a thorough discussion on the topic.

Finally, the annotators involved in the study were paid more than the minimum wage per hour.

Acknowledgments

Nedjma Ousidhoum would like to thank Abderrahmane Samir Lazouni, Lyes Taher Khalfi, Manel Amarouche, Narimane Zahra Boumezrag, Noufel Bouslama, Raouf Ousidhoum, Sarah Arab, Wassil Adel Merzouk, Yanis Rabai, and another annotator who would like to stay anonymous for their work and insightful comments.

Jan Philip Wahle, and Terry Ruas, Bela Gipp, Florian Wunderlich were partially supported by the Lower Saxony Ministry of Science and Culture and the VW Foundation.

Yi Zhou would like to thank Gaifan Zhang, Bing Xiao, and Rui Qin for their help with the annotations and for providing feedback.

Alexander Panchenko would like to thank Nikolay Ivanov, Artem Vazhentsev, Mikhail Salnikov, Maria Marina, Vitaliy Protasov, Sergey Pletenev, Daniil Moskovskiy, Vasiliy Konovalov, Elisey Rykov, and Dmitry Iarosh for their help with the annotation, and was partially funded by AIRI.

Rahmad Mahendra acknowledges the funding by

Hibah Riset Internal Faculty of Computer Science, Universitas Indonesia under contract number: NKB-13/UN2.F11.D/HKP.05.00/2024 for supporting the annotation of the Indonesian, Javanese, and Sundanese datasets.

References

Magda B Arnold. 1960. Emotion and personality. vol. i. psychological aspects.

L.F. Barrett. 2017. *How Emotions are Made: The Secret Life of the Brain*. Expert Thinking Series. Macmillan.

Lisa Feldman Barrett. 2016. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2021. FEEL-IT: Emotion and sentiment classification for the Italian language. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 76–83, Online. Association for Computational Linguistics.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2022. Xlm-emo: Multilingual emotion prediction in social media text. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 195–203.

Kateryna Bobrovnyk. 2019. Automated building and analysis of ukrainian twitter corpus for toxic text detection. In *COLINS 2019. Volume II: Workshop*.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alexandra Ciobotaru and Liviu P. Dinu. 2021. RED: A novel dataset for Romanian emotion detection from tweets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 291–300, Held Online. INCOMA Ltd.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of* *the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Paul Ekman. 1992. Are there basic emotions?

PC Ellsworth. 2013. Appraisal theory: old and new questions. emot. rev. 5, 125–131.

Nico H Frijda. 1986. The emotions. *Studies in Emotion and Social Interaction*.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

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.

MD Asif Iqbal, Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H Sarker. 2022. Bemoc: A corpus for identifying emotion in bengali texts. *SN Computer Science*, 3(2):135.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego, California.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, 240:108112.

Richard S Lazarus. 1991. *Emotion and adaptation*, volume 557. Oxford University Press.

Youssef Mohamed, Runjia Li, Ibrahim Said Ahmad, Kilichbek Haydarov, Philip Torr, Kenneth Church, and Mohamed Elhoseiny. 2024. No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20939–20962, Miami, Florida, USA. Association for Computational Linguistics.

Saif Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836, Dubrovnik, Croatia. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018a. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018b. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saif M. Mohammad. 2016. 9 - sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Herbert L. Meiselman, editor, *Emotion Measurement*, pages 201–237. Woodhead Publishing.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Preprint*, arXiv:2109.08256.

Saif M. Mohammad. 2024. WorryWords: Norms of anxiety association for over 44k English words. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16261–16278, Miami, Florida, USA. Association for Computational Linguistics.

Agnes Moors, Phoebe C Ellsworth, Klaus R Scherer, and Nico H Frijda. 2013. Appraisal theories of emotion: State of the art and future development. *Emotion review*, 5(2):119–124.

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, Felermino 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. 2023a. 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.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ayele, Saif M. Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023b. SemEval-2023 task 12: Sentiment analysis for african languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. Xed: A multilingual dataset for sentiment analysis and emotion detection. *arXiv preprint arXiv:2011.01612*.

Andrew Ortony, Gerald L Clore, and Allan Collins. 2022. *The cognitive structure of emotions*. Cambridge university press.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

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.

Robert Plutchik. 1980. Chapter 1 - a general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press.

Ira J Roseman. 2013. Appraisal in the emotion system: Coherence in strategies for coping. *Emotion Review*, 5(2):141–149.

Alieh Hajizadeh Saffar, Tiffany Katharine Mann, and Bahadorreza Ofoghi. 2023. Textual emotion detection in health: Advances and applications. *Journal of Biomedical Informatics*, 137:104258.

Mei Silviana Saputri, Rahmad Mahendra, and Mirna Adriani. 2018. Emotion classification on indonesian twitter dataset. In *2018 International Conference on Asian Language Processing (IALP)*, pages 90–95.

Klaus R Scherer. 2009. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and emotion*, 23(7):1307–1351.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Krishnapriya Vishnubhotla, Daniela Teodorescu, Mallory J Feldman, Kristen Lindquist, and Saif M. Mohammad. 2024. Emotion granularity from text: An aggregate-level indicator of mental health. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19168–19185, Miami, Florida, USA. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

A Annotation

A.1 Annotation Guidelines and Definitions

This is a guide for annotating text for emotion classification. The purpose of this study is to analyze the emotions expressed in a text. It is important to note that emotions can often be inferred even if they are not explicitly stated.

Task The task involves classifying text into predefined emotion categories. The annotated dataset will be used for training emotion classification models and studying how emotions are conveyed through language.

Emotion Categories We categorize emotions into the following seven classes:

Joy

- Definition: Expressions of happiness, pleasure, or contentment.
- Example: *"I just passed my exams!"*

Sadness

- Definition: Expressions of unhappiness, sorrow, or disappointment.
- Example: *"I miss my family so much. It's been a tough year."*

Anger

- Definition: Expressions of frustration, irritation, or rage.
- Example:*"Why is the internet so slow today?!"*

Fear

- Definition: Expressions of anxiety, apprehension, or dread.
- Example: *"There's a huge storm coming our way. I hope everyone stays safe."*

Surprise

- Definition: Expressions of astonishment or unexpected events.
- Example: *"I can't believe he just proposed to me!"*

Disgust

- Definition: A reaction to something offensive or unpleasant.
- Examples: *"That video was sickening to watch."*

Neutral

- Definition: Texts that do not express any of the above emotions.
- Example: *"The weather today is sunny with a chance of rain."*

Note: Factual statements can indicate an emotional state without explicitly stating it. For example:

- *"An earthquake today killed hundreds of people in my home town."*

Surprise differs from joy in that it represents an unexpected event, which may or may not be associated with happiness.

Emotion Description Categories The following list provides a broader categorization of emotions by including synonyms and related emotional states.

Anger

- Includes: *irritated, annoyed, aggravated, indignant, resentful, offended, exasperated, livid, irate*, etc.

Sadness

- Includes: *melancholic, despondent, gloomy, heartbroken, longing, mourning, dejected, downcast, disheartened, dismayed*, etc.

Fear

- Includes: *frightened, alarmed, apprehensive, intimidated, panicky, wary, dreadful, shaken*, etc.

Happiness

- Includes: *joyful, elated, content, cheerful, blissful, delighted, gleeful, satisfied, ecstatic, upbeat, pleased*, etc.

Surprise

- Includes: *taken aback, bewildered, astonished, amazed, startled, stunned, shocked, dumbstruck, confounded, stupefied*, etc.

Joy

- Includes: *happiness, delight, elation, pleasure, excitement, cheerfulness, bliss, euphoria, contentment, jubilation*.

A.1.1 Emotion Intensity

After selecting the emotion category, annotators were further asked to select the intensity label, which could be: 0: No Emotion, 1 - Slight Emotion, 2: Moderate Emotion and 3: High Emotion. The following examples illustrate different levels of emotion intensity.

Anger

- No Anger: *"I walked through the empty streets, the quiet hum of the city like a distant whisper."*

- Slight Anger: *"The buzz of voices around me blended into a monotonous drone, failing to distract from the pang of annoyance at the delay."*

- High Anger: *"When his friend's brother knocked on the door, he was greeted with a shotgun blast through the door, which left him dead at the doorstep."*

A.2 Pilot Annotation

We run a pilot annotation on different languages to further refine our guidelines. This has mainly led to further clarifications related to the labeling process. For instance, the annotators were reminded that they should select all the labels that apply for a given text snippet, and that one label can encompass more than one specific emotion (e.g., in `arq`, we explained that a complex perceived emotion such as bitterness or jealousy might involve both anger and sadness).

A.3 Formula for Determining Final Labels

Aggregating emotion labels Aggregating emotion labels can be formally expressed as:

$$L_{\text{final}} = \begin{cases} 1, & \text{if Count}(1,2,3) \geq 2 \text{ and AvgScore} > T, \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{Count}(1,2,3) = \sum_{i=1}^{N} \nVdash(A_i \in \{1,2,3\}),$$

$$\text{AvgScore} = \frac{1}{N}\sum_{i=1}^{N} A_i,$$

Where:

- A_i is the rating provided by annotator i.

- N is the total number of annotators.

- $\nVdash(A_i \in \{1,2,3\})$ Membership function that returns 1 if $A_i \in \{1,2,3\}$, and 0 otherwise.

- T is the threshold for the average score, which we set as $T = 0.5$

Aggregating intensity Aggregating intensity can be formally expressed as:

$$\text{AvgScore} = \frac{\sum_{i=1}^{N} A_i}{N},$$

$$L_{\text{final}} = \begin{cases} 0, & \text{if } 0 \leq \text{AvgScore} < 1, \\ 1, & \text{if } 1 \leq \text{AvgScore} < 2, \\ 2, & \text{if } 2 \leq \text{AvgScore} < 3, \\ 3, & \text{if AvgScore} = 3. \end{cases}$$

Where:

- A_i is the intensity score provided by annotator i, where $A_i \in \{0,1,2,3\}$.

- N is the total number of annotators.

B SCHMP Calculation

The computation of SHCMP involves the following steps:

1. Random Splitting with Tie-Breaking The dataset of N annotated items is randomly divided into two equal subsets, A_1 and A_2. For datasets with an odd number of annotations, probabilistic tie-breaking is applied to ensure balanced splits.

2. Class Assignment For each item $x_i \, (i = 1, 2, \ldots, N)$:

- Assign x_i a score based on its annotations in A_1 and A_2.
- Let $C_1(x_i)$ and $C_2(x_i)$ denote the class of x_i derived from A_1 and A_2, respectively.

3. Class Binning To manage continuous scores, divide the range of possible scores $[-3, 3]$ into equal-sized bins, where the bin size b is determined as:

$$b = \frac{6}{\#\text{Bins}}.$$

Scores from A_1 and A_2 are then assigned to their respective bins, denoted as c_1 and c_2.

4. Match Calculation Define a match indicator $M(x_i)$ to evaluate consistency for each item:

$$M(x_i) = \begin{cases} 1, & \text{if } |c_1 - c_2| < 1, \\ 0, & \text{otherwise.} \end{cases}$$

This ensures that items are considered consistent if their scores fall into the same bin or adjacent bins.

5. Proportion of Matches Compute the total number of matches, N_{match}, across all items:

$$N_{\text{match}} = \sum_{i=1}^{N} M(x_i).$$

6. SHCMP Computation The SHCMP score is calculated as the proportion of matches, expressed as a percentage:

$$\text{SHCMP (\%)} = \frac{N_{\text{match}}}{N} \times 100.$$

7. Averaging We repeat the process k times with different random splits and compute the average SHCMP score:

$$\text{SHCMP}_{\text{final}} = \frac{1}{k} \sum_{j=1}^{k} \text{SHCMP}_j,$$

where SHCMP_j is the SHCMP score from the j-th split.

C Experimental Settings

For LLMs, we used the default parameters from HuggingFace except for temperature which we set to 0 for deterministic output and topk is set to 1. Only for the topk ablations in which topk > 1 in Figure 5c, we set temperature to 0.7. We ask all LLMs to perform CoT. We trained on the train set for 2 epochs with a learning rate of 1e-5 and and evaluated on test set. For MLMs experiments, we trained on the training set for 2 epochs with a learning rate of 1e-5 and evaluated on the test set.

Language	Train Set (%)			Development Set (%)			Test Set (%)		
	Single	Multi	Neutral	Single	Multi	Neutral	Single	Multi	Neutral
chn	54.00	23.74	22.26	53.60	23.58	22.82	53.90	24.30	21.80
sun	58.94	36.18	4.88	59.09	36.26	4.65	59.40	36.07	4.54
afr	47.79	6.69	45.52	56.14	7.86	36.01	37.39	10.35	52.26
swe	43.16	16.60	40.24	46.30	20.37	33.33	42.76	18.81	38.43
swa	41.67	3.33	55.00	45.78	3.56	50.66	46.26	3.81	49.93
esp	61.02	38.98	0.00	65.22	34.78	0.00	65.14	34.86	0.00
arq	28.53	50.05	9.42	28.57	50.00	10.71	27.95	44.76	8.35
ptbr	52.11	13.80	34.09	61.06	11.82	27.12	52.68	13.59	33.73
ptmz	52.00	0.44	47.56	50.92	0.37	48.71	53.03	0.51	46.45
ukr	44.77	2.24	52.99	47.24	2.36	50.39	45.23	1.79	52.98
mar	67.69	8.56	23.75	68.57	7.62	23.81	68.94	9.33	21.73
rus	64.63	11.08	24.29	66.35	12.23	21.42	66.91	12.89	20.20
ibo	72.44	3.63	23.93	61.12	10.91	27.97	73.61	3.97	22.42
amh	50.82	27.68	21.50	56.13	30.31	16.56	48.50	24.67	26.83
deu	41.78	34.05	24.17	41.84	35.19	22.97	41.23	32.10	26.66
vmw	52.80	0.45	46.75	53.49	0.39	46.12	53.46	0.52	46.32
pcm	55.00	40.46	4.54	50.00	36.63	4.37	51.57	38.08	4.35
eng	38.64	47.02	14.34	34.07	42.22	9.70	38.58	48.76	10.34
hin	66.35	10.80	22.85	60.40	7.92	31.68	77.31	5.66	13.92
tat	81.48	0.00	18.52	84.00	0.00	16.00	85.71	0.00	14.29

Table 4: Percentage distribution of SingleLabel, MultiLabel, and NeutralLabel for the Train, Development, and Test Sets.

	Monolingual Multi-Label Classification				
Lang.	LaBSE	RemBERT	XLM-R	mBERT	mDeBERTa
afr	30.76	37.14	10.82	25.87	16.66
arq	45.46	41.41	31.98	41.75	29.68
ary	45.81	47.16	40.66	36.87	38.00
chn	53.47	53.08	58.48	49.61	44.47
deu	55.02	64.23	55.37	46.78	44.09
eng	64.24	70.83	67.30	58.26	58.94
esp	72.88	77.44	29.85	54.41	60.17
hau	58.49	59.55	36.95	47.33	48.59
hin	75.25	85.51	33.71	54.11	54.34
ibo	45.90	47.90	18.36	37.23	31.92
ind	–	–	–	–	–
jav	–	–	–	–	–
kin	50.64	46.29	32.93	35.61	38.00
mar	80.76	82.20	78.95	60.01	66.01
pcm	51.30	55.50	52.03	48.42	46.21
ptbr	42.60	42.57	15.40	32.05	24.08
ptmz	36.95	45.91	30.72	14.81	21.89
ron	69.79	76.23	65.21	61.50	60.60
rus	75.62	83.77	78.76	61.81	54.79
sun	36.93	37.31	19.66	27.88	21.65
swa	27.53	22.65	22.71	22.99	22.84
swe	49.23	51.98	34.63	44.24	40.90
tat	57.71	53.94	26.48	43.49	35.02
ukr	50.07	53.45	17.77	31.74	28.55
vmw	21.13	12.14	9.92	10.28	11.13
xho	–	–	–	–	–
yor	32.55	9.22	11.94	21.03	17.88
zul	–	–	–	–	–

Table 5: **Average F1-Macro for monolingual multi-label emotion classification.** Each model is trained and evaluated within the same language.

Prompt Version	**Prompt Text**
Prompt v1	Evaluate whether the following text conveys the emotion of {{EMOTION}}. Think step by step before you answer. Finish your response with ’Therefore, my answer is ’ followed by ’yes’ or ’no’: {{INPUT}}
Prompt v2	Analyze the text below for the presence of {{EMOTION}}. Explain your reasoning briefly and conclude with ’Answer:’ followed by either ’yes’ or ’no’. {{INPUT}}
Prompt v3	Examine the following text to determine whether {{EMOTION}} is present. Provide a concise explanation for your assessment and end with ’Answer:’ followed by either ’yes’ or ’no’. {{INPUT}}

Table 6: The prompt variants used for ablation of Track A.

Track A: Example Few-Shot Prompt

### Task: ###
Analyze the text below for the presence of anger.
Explain your reasoning briefly and conclude with ’Answer:’ followed by either ’yes’ or ’no’.

### Examples: ###
Example 1:
Input: ”’When I answered the phone, my heart beat extremely fast... I was very nervous!”’
Answer: no

Example 2:
Input: ”’I’ll never forget how businesslike and calm the Israeli guy was.”’
Answer: no

Example 3:
Input: ”’I wake up, my eyes fluttering open to a shield of darkness.”’
Answer: no

Example 4:
Input: ”’I lay in a large bed, the sheets and quilt pulled up to my chin, and the curtains were drawn to keep out the light.”’
Answer: no

Example 5:
Input: ”’Either way that idiot is gone.”’
Answer: yes

Example 6:
Input: ”’Seriously... did I really just shut my finger in the car door.”’
Answer: yes

Example 7:
Input: ”’I was really uncomfortable because I was sitting behind my dad and there isn’t enough room for my legs.”’
Answer: yes

Example 8:
Input: ”’He damn disturb plz, cover my head with a shirt that a customer which have body odour just tried on!!”’
Answer: yes

### Your Turn: ###
Input: ”’/ o So today I went in for a new exam with Dr. Polvi today, I had to file new paperwork for the automobile accident case which is being done differently than the scoliosis stuff. So he comes in and starts talking about insurance stuff and how this looks bad since I was getting treatment on my neck and stuff already blah blah.”’

Figure 7: Example of the few-shot prompt template for assessing anger in Track A.

Track B: Example Few-Shot Prompt

### Task: ###

In this task, you will assess the level of anger in a given text (0 = none, 1 = low, 2 = medium, 3 = high). Summarize your reasoning and conclude with ’Answer:’ followed by the correct number.

### Examples: ###

Example 1:
Input: ”’I try extremely hard to keep my details hidden. It was nice to know that what I had given people to know was pleasant, but I couldn’t deny the knot that was still in my stomach.”’
Answer: 0

Example 2:
Input: ”’I knew we were almost there when my midwife’s voice got more excited and Joey leaned in real close and said into my ear, “ Don’t stop pushing! ” ”’
Answer: 0

Example 3:
Input: ”’One ended up going to prison.”’
Answer: 1

Example 4:
Input: ”’Not to mention noisy.”’
Answer: 1

Example 5:
Input: ”’" but Urban Dictionary confirmed Spook is indeed a racial slur.”’
Answer: 2

Example 6:
Input: ”’And..at his funeral, they fired him!”’
Answer: 2

Example 7:
Input: ”’I ended up metaphorically throwing my hands in the air in disgust and just cancelling my account altogether.”’
Answer: 3

Example 8:
Input: ”’He would manipulate me into it and I was extremely upset.”’
Answer: 3

### Your Turn: ###

Input: ”’So today I went in for a new exam with Dr. Polvi today, I had to file new paperwork for the automobile accident case which is being done differently than the scoliosis stuff. So he comes in and starts talking about insurance stuff and how this looks bad since I was getting treatment on my neck and stuff already blah blah.”’

Figure 8: Example of the few-shot prompt template for assessing anger in Track B.