

Emo Pillars π^{emo} : Knowledge Distillation to Support Fine-Grained Context-Aware and Context-Less Emotion Classification

Alexander Shvets

NLP Group, Pompeu Fabra University, Barcelona, Spain

Language Technologies Unit, Barcelona Supercomputing Center, Spain

alexander.shvets@upf.edu, aleksandr.shvets@bsc.es

Abstract

Most datasets for sentiment analysis lack context in which an opinion was expressed, often crucial for emotion understanding, and are mainly limited by a few emotion categories. Foundation large language models (LLMs) like GPT-4 suffer from over-predicting emotions and are too resource-intensive. We design an LLM-based data synthesis pipeline and leverage a large model, Mistral-7b, for the generation of training examples for more accessible, lightweight BERT-type encoder models. We focus on enlarging the semantic diversity of examples and propose grounding the generation into a corpus of narratives to produce non-repetitive story-character-centered utterances with unique contexts over 28 emotion classes. By running 700K inferences in 450 GPU hours, we contribute with the dataset of 100K contextual and also 300K context-less examples to cover both scenarios. We use it for fine-tuning pre-trained encoders, which results in several Emo Pillars π^{emo} models. We show that π^{emo} models are highly adaptive to new domains when tuned to specific tasks such as GoEmotions, ISEAR, IEMOCAP, and EmoContext, reaching the SOTA performance on the first three. We also validate our dataset, conducting statistical analysis and human evaluation, and confirm the success of our measures in utterance diversification (although less for the *neutral* class) and context personalization, while pointing out the need for improved handling of out-of-taxonomy labels within the pipeline.

1 Introduction

Manual tagging of texts with emotion categories is a very burdensome process. It is overwhelmed with a multitude of patternless ways to express the same emotion. At the same time, similar utterances may refer to a different set of emotions, as they are often context- and personality-dependent, which hinders establishing a basis for consistent annotation, especially within a fine-grained manifold (Devillers

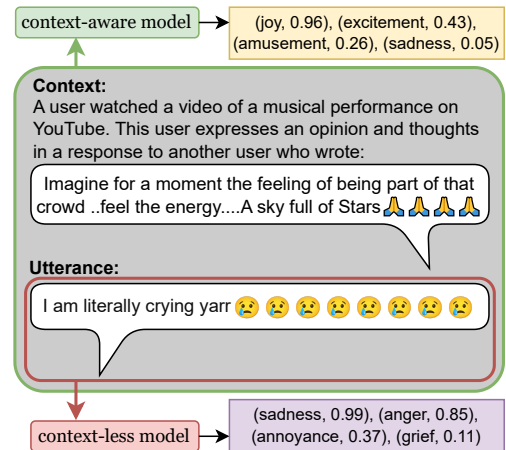


Figure 1: Difference in context-less (context cannot be taken into account) and context-aware (context helps) emotion classification. Context-less models detect emotions in the entire input (including context if provided), while context-aware models can grasp the input structure and extract emotions only from the utterance.

et al., 2005). Most datasets in sentiment analysis contain overgeneralized coarse-grained schemes with rare exceptions in context-less settings that would consider a fair number of labels (Tu et al., 2022; Lykousas et al., 2019; Demszky et al., 2020). However, without context, the task is highly subjective, and the interpretation of a situation depends solely on the annotators’ priors, leading to large disagreements (Park et al., 2021).

Figure 1 depicts the importance of context in automatic emotion understanding. Models trained in a regular context-less setup would either ignore the provided situation or extract emotions from the entire input, not being able to distinguish its parts. Foundation large language models that solve various contextual tasks without task-specific tuning (LLMs), e.g. GPT-{3.5,4} (OpenAI, 2024), can grasp the compound structure of the input, but they understand emotions poorly (Wang et al., 2024; Chochlakis et al., 2024; Sabour et al., 2024).

Synthetic data generation using LLMs has be-

come a common practice in addressing the data scarcity problem. However, it possesses many limitations, with the lack of diversity of the generated data being one of the main issues, especially for high-subjectivity tasks (Li et al., 2023). Models tend to generalize too much and produce rather similar items. This requires looking for techniques to make outputs more diverse, such as special “diversity prompts”, incorporation of real-world data examples within a few-shot setting, and varying the temperature parameter, which can lead to less realistic semantics (Li et al., 2023; Kok-Shun et al., 2023; Jin et al., 2024). The problem is even more acute for massive inference (Yang et al., 2023).

We propose to use a combination of measures to enhance the semantic diversity of LLMs’ outcomes when extracting knowledge to create a large emotion classification dataset: (i) grounding examples into a variety of story texts from topic-rich corpora rather than using generic instructions, (ii) evoking the model to consider each text many times from the perspective of different characters and guess their thoughts and emotions that varied throughout the narrative, and (iii) requesting the model to generate several thoughts/utterances at once while reducing the chances of word repetition at the level of model parameters. In addition, to mitigate the issue of subjectivity, we propose to prompt the model to choose multiple additional “soft” co-occurring labels for each example rather than stick to a single “hard” one and create utterance-centered summaries of ground texts that would decrease the number of interpretations. Finally, to enhance the robustness of classifiers, we propose to make emotions less explicit in an utterance so that reliance on the context would become a key to understanding them.

To unite these measures, we design an LLM-based data synthesis pipeline that covers many emotions and topics, relates well contexts to utterances, and maintains realistic semantics. We produce and leverage the data to fine-tune mid-sized encoder-based models to *support* context-aware and context-less scenarios encountered in real-world applications. The resulting *Emo Pillars* $\frac{\text{emo}}{\pi}$ models that are not prone to hallucinations (as they are non-autoregressive) and require less computing than LLMs at inference, reach SOTA performance on several tasks through transfer learning. We release our code¹, dataset², and models³ openly.

¹<https://github.com/alex-shvets/EmoPillars>

²huggingface.co/datasets/alex-shvets/EmoPillars

³huggingface.co/collections/alex-shvets/EmoPillars

2 Related Work

Despite the very advanced abilities of LLMs they still can hardly grasp sentiments. Wang et al. (2024) provided a good error analysis showing three main issues when LLMs are applied as-is: hallucination, over-labelling, and over-interpretation. Chochlakis et al. (2024) added that few-shot learning does not improve the outcome a lot as LLMs are unable to fully integrate information from subjective demonstrations that contrast their strong task priors. Instead, they tend to ossify their predictions; the larger the model, the stronger the effect is observed.

Medium-sized language models, like RoBERTa (Liu et al., 2019), remain a better option (Alvarez-Gonzalez et al., 2021; Park et al., 2021; Zanwar et al., 2022; Cortiz, 2022; Kok-Shun et al., 2023), although they require more training examples per emotion than are usually available in existing datasets. Data augmentation using LLMs has become a common practice in enlarging minor classes. However, it is limited by creating derivatives of the dataset examples through paraphrasing and inspirational generation based on them (Wang et al., 2024), which does not bring much content diversity and may not always align with the expected sentiment (Woźniak and Kocoń, 2023). Another type of data enhancement for a given task, e.g. as in Koufakou et al. (2023), consists in integrating an external dataset such as GoEmotions (Demszky et al., 2020), which is so far the largest third-person annotated fine-grained sentiment dataset of 58K Reddit comments over 28 labels. As it is highly unbalanced, it is either taken only partially with dominant labels (Park et al., 2021; Zanwar et al., 2022) or with pre-augmentation, as e.g. in Kok-Shun et al. (2023) using GPT-{3.5,4} with a simple 5-shot prompt asking to generate 20 records for a provided emotion. Moreover, comments of GoEmotions can be perceived ambiguously in the absence of context, and therefore, alternative labels can be acceptable (Kocoń et al., 2023). The latter happens even with first-party annotated comments (Malko et al., 2021). To reduce the number of possible interpretations in GoEmotions, Yang et al. (2023) proposed to extend the comments by prompting GPT-{3.5,4} to add 1-2 sentences at the end so that they clearly convey the ground truth labels. However, their analysis showed reduced vocabulary diversity as the same 3 words are reused in 20-30% cases within an emotion. This brings a risk of over-tuning the models to limited patterns.

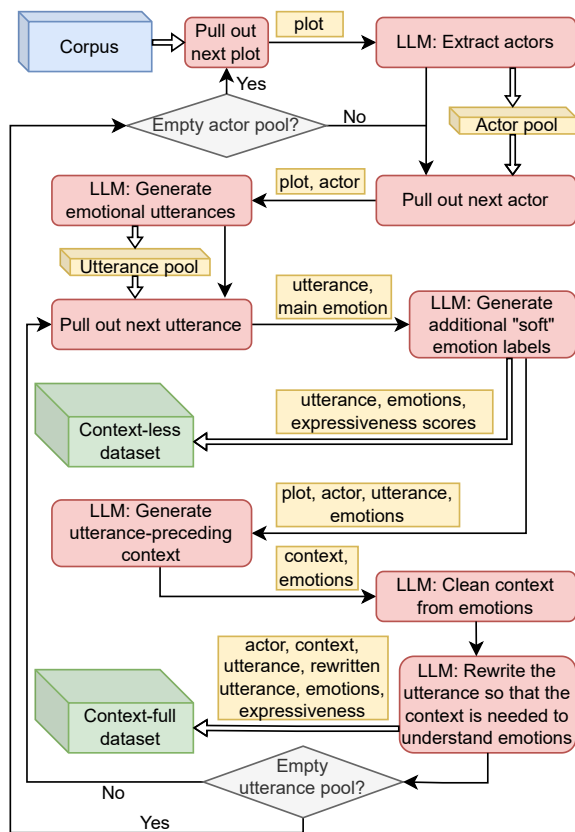


Figure 2: Our pipeline for the generation of a dataset for a multi-label context-aware (-less) emotion recognition.

3 Data Synthesis Pipeline

Figure 2 depicts our procedure of example generation. It is described in the subsequent subsections, while the implementation details are in Section 4. We perform prompting in several successive steps to let the model focus on a single task at a time, thus having fewer constraints on generating the next tokens. Since the structure of answers has minor variations for a fixed prompt, we can parse them with simple regex-based extractors. Table 15 in Appendix B lists the designed prompts and corresponding outcomes.

3.1 Content-rich instructions

The uniqueness of instructions is reached by iterating over texts in a corpus of narratives (e.g., story plots or news items) and different characters (actors). The actors are identified using an LLM with a simple prompt, as we are not aiming for precise extraction. We iterate over all the found actors to increase the coverage of many diverse emotions and alternative stances raised in the text.

3.2 Multiple example generation

We request the model to produce several utterances of the same actor within a single inference and also cover various emotional classes: this gives a model a higher chance of avoiding similar utterances for emotions that can co-occur (i.e., be expressed within the same utterance) and resulting in more contrasting examples where the primary emotion would be presented more expressively. We provide definitions of emotions in the prompt to reduce ambiguity for related labels such as, for example, *sadness* and *disappointment*. Along with emotional utterances, we request two *neutral* ones at the same inference to only slightly anticipate the real-world prevalence of neutral sentiment.

3.3 Soft labelling

At this step, we assign multiple emotion labels for an utterance. We eliminate the plot from the prompt so that an LLM does not attend to the actor’s emotions in other moments of the story and thus has less information to hallucinate. Instead, we provide the primary emotion for which the utterance was generated. This allows for mitigating the problem of ambiguous sentiment for short utterances. We request the labels to be soft, i.e., with an assessed expressiveness level from 0 to 1 with a step of 0.1. These scores are independent and thus do not sum up to 1. We select labels scored above an established threshold of 0.3 for further steps.

3.4 Context generation and cleaning

This step is to reconstruct the situation in which the character arrives to pronounce the utterance. The challenging task is to formulate a prompt so that the LLM does not summarize the entire plot but centers the context around the character and ends it before the moment the utterance is expressed, not revealing the emotions. We found out that the models are sensitive to slight changes in wording in this case. For example, before adding “Be as concise as possible.”, the output was twice as long and often included information describing post events and emotions⁴. To ensure the absence of explicit emotions in contexts, we complemented the pipeline with a context-cleaning step that removes emotive clauses and adjusts affected sentences.

⁴For example, some context would include a sentence “He begins to see her as more than just a replicant and develops feelings of caring and protection towards her.”

3.5 Context importance upscale

Our context generation aims to provide more clarity to utterances. However, in case an utterance is self-explanatory, i.e., it contains enough signals to interpret emotions, a model may not learn to attend to the information in contexts during the training phase. Then, at prediction time, the attention weights will be kept low for the context tokens even when they are necessary for deriving correct labels. To ensure the importance of the contexts in the dataset, while observing that our generated utterances are rather detailed, we added a step of their rewriting. We specify in the prompt that the context shall become crucial for the emotion understanding by reducing the explicitness of emotions in the utterance. In the following example, the updated utterance drops the markers “scared” and “safely” that were emphasizing the emotions of *fear* and *caring* correspondingly:

Original: “I’m really scared right now. I don’t know what to do. I need to get this plane down safely.”

Rewritten: “I’m not sure what to do. I need to land this plane.”

If the original utterance expresses emotions mildly, the change is minor and the meaning is kept intact.

4 Dataset Generation

The seed corpus we use is compiled of 113K synopses of movies, books, and TV shows from English Wikipedia⁵. According to Papalampidi et al. (2022), the stories of this corpus contain intervening events and non-linearities, as well as many characters with elaborate and diverse attributes.

We limit the emotion taxonomy to the 28 GoEmotions categories using definitions from its paper (provided in Appendix A), for better comparability.

The choice of an LLM fell on the Mistral model⁶, a leading model with a large context window that allows for the reuse of its outcomes and can be cost-time-effectively used for massive inference (Jiang et al., 2023). As the model sometimes hallucinates and generates emotions out of the provided list, we perform manual mapping of frequent labels to our categories (e.g., *anxiety* to *nervousness*, *indignation* to *anger*, *hope* to *optimism*, *happiness* to *joy*),

⁵<https://github.com/markriedl/WikiPlots>

⁶[mistralai/Mistral-7B-Instruct-v0.2](https://mistralai.com/). Details on prompting are in Appendix B. We also tested prompts with GPT-3.5 to keep them generic. Cf. also [Limitations](#) for the model choice.

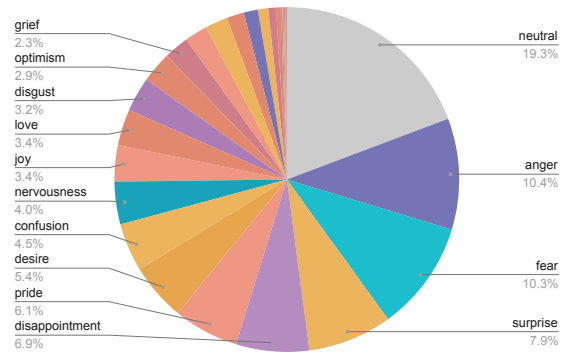


Figure 3: Distribution of primary emotions.

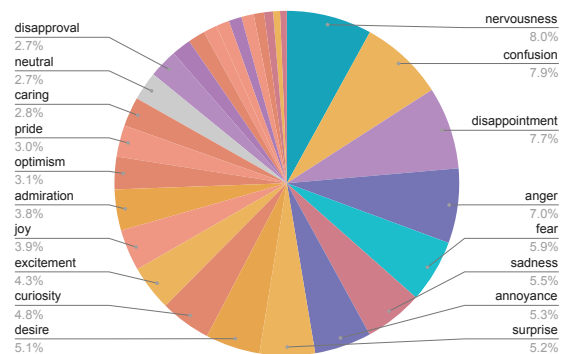


Figure 4: Distribution of soft emotional labels in the dataset (after filtering by the expressiveness level).

while removing labels for which we did not find a close match (such as *calm* and *focus*). We also store the model’s reasoning with emotion explications.

We ran our data synthesis pipeline for 450 GPU hours of NVIDIA H100 iterating over 2000 plots to generate 300K examples for a context-less dataset, a third of which was used for the final pipeline steps, resulting in a context-full dataset of 100K examples (cf. the note on a carbon footprint and scalability in Appendix C). On average, 15 actors were extracted from a plot (*std* of 9.44). The distribution of 300K primary emotions is shown in Figure 3, while Figure 4 shows the improved class balance after extending them to 1M soft labels for which the expressiveness level was of at least 0.3 (1 to 5 labels per example, with a *mean* of 3.17 and *std* of 0.97). We create 80-10-10 and 90-5-5⁷ splits for the train, dev, and test sets as shown in Figure 5.

5 Experiments

We first perform regular sentiment classification when only an utterance is provided as input. With

⁷The second dataset is three times smaller. We sampled 90% for the training to have more examples in absolute value.

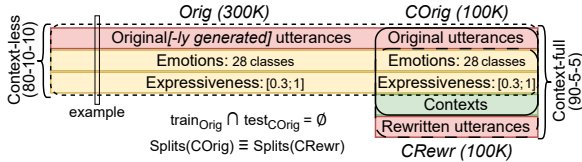


Figure 5: Dataset splits. *Orig* – context-less examples, *COrig* – context-full examples, *CReWr* – the same context-full examples with rewritten utterances.

this, we check that the LLM managed to include emotion-related information in the *original* utterances enough for smaller models to grasp the nuances in various categories and perform the task well. Then, we run experiments using generated contexts paired with either *original* or *rewritten* utterances to investigate whether our contexts, and hence the models trained using them, bring added value for contextual emotion detection tasks. All the base models were trained for multi-label classification with a sigmoid activation function for each category⁸ and binary cross-entropy loss⁹ using the *transformers* library with the default AdamW optimizer, and initial *learning_rate* of 2e-5. We did not activate early stopping and worked with the final checkpoints for all the tasks. The details of their domain fine-tuning are provided in the task-related subsections. The scores are macro-averaged as micro-averaging over-promise results in all tasks.

5.1 Context-less emotion recognition

5.1.1 Emo Pillars synthetic test sets

Our first experiment was directed to study to what extent encoder-decoder transformer models may follow the labels derived using Mistral’s inference. We trained two RoBERTa models¹⁰ for sequence classification, one on *original* utterances from the context-less dataset and another on *rewritten* ones from the context-full dataset, excluding contexts. We set *num_train_epochs* to 10, *max_seq_length* to 128, and *batch_size* to 64. To contrast the utterance types, we evaluate each model on both of them. We also trained one version of BERT¹¹ (Devlin et al., 2019) and SentenceBERT¹² (SBERT) (Reimers and Gurevych, 2019) to further fine-tune them on downstream tasks and compare with previous-work models of the same pre-trained architectures. The results are in Table 1.

⁸Cf. Appendix D for lower boundary choice at evaluation.

⁹<https://pytorch.org/docs/.../torch.nn.BCELoss.html>

¹⁰FacebookAI/roberta-large

¹¹google-bert/bert-large-uncased, used in 5.1.2.

¹²sentence-transformers/paraphrase-distilroberta, in 5.1.4.

Model	Eval set	P	R	F ₁
$\overset{\text{emo}}{\pi}$ -BERT _{Orig}	test _{Orig}	0.81	0.79	0.80
$\overset{\text{emo}}{\pi}$ -SBERT _{Orig}	test _{Orig}	0.78	0.80	0.79
$\overset{\text{emo}}{\pi}$ -RoBERTa _{Orig}	test _{Orig}	0.82	0.79	0.81
$\overset{\text{emo}}{\pi}$ -RoBERTa _{ReWr}	test _{Orig}	0.70	0.78	0.74
$\overset{\text{emo}}{\pi}$ -RoBERTa _{Orig}	test _{COrig}	0.82	0.80	0.81
$\overset{\text{emo}}{\pi}$ -RoBERTa _{ReWr}	test _{COrig}	0.71	0.78	0.74
$\overset{\text{emo}}{\pi}$ -RoBERTa _{Orig}	test _{CReWr}	0.68	0.63	0.65
$\overset{\text{emo}}{\pi}$ -RoBERTa _{ReWr}	test _{CReWr}	0.73	0.66	0.69

Table 1: Context-less intra-dataset evaluation. Subscripts show the type of utterances, while “C” signifies examples of the context-full test set (contexts excluded).

We observe that the models trained on non-restricted *original* utterances reach high scores when applied to test examples from the same set (F_1 of 0.79-0.81). This indicates that there are enough signals in the generated texts to identify emotions even without context. However, such models perform poorer on less emotive material (i.e., *rewritten* utterances) than the model trained for this purpose (F_1 of 0.65 vs. 0.69). On the other hand, the latter model performs worse on longer texts (F_1 of 0.74 vs. 0.81) as it learns to predict emotions beyond explicitly mentioned, and this results in over-labelling and hence low precision.

5.1.2 GoEmotions

GoEmotions is a multi-label categorical dataset of 58K English Reddit comments with 28 emotion labels, including *neutral* class (Demszky et al., 2020). There are some drawbacks in its quality such as rather low inter-annotator agreement and large disparity in terms of emotion frequencies (Park et al., 2021). However, it is still a very valuable source for preliminary evaluation to get a reference point in highly fine-grained emotion classification. We use the original pre-splits for the train, validation, and test sets of the dataset.

We show the performance of the base models applied as is in the upper part of Table 2, while the bottom part provides models fine-tuned specifically on this dataset. We ran fine-tuning of our pre-trained $\overset{\text{emo}}{\pi}$ models for 3 epochs with a batch size of 16. Even though our models operate within the same set of categories, fine-tuning allows adapting them to the domain, i.e., to the writing style pertinent to Reddit and specific topics. To the best of our knowledge, the F_1 -score of 0.55 (*std* of 0.007 over 3 runs) makes our model SOTA on this task¹³.

¹³We provide per-class evaluation scores in Appendix E.

Model	P	R	F ₁
GPT4 (Wang et al., 2024)	0.10	0.17	0.13
GPT4 (Kok-Shun et al., 2023)	-	-	0.22
$\frac{\text{emo}}{\pi}$ -BERT _{Orig}	0.26	0.42	0.28
$\frac{\text{emo}}{\pi}$ -RoBERTa _{Orig}	0.25	0.45	0.28
$\frac{\text{emo}}{\pi}$ -RoBERTa _{Rewr}	0.22	0.33	0.22
BERT-based (Demszky et al., 2020)	0.40	0.63	0.46
BERT-based (Alvarez-Gonzalez et al., 2021) ¹⁴	-	-	0.48
RoBERTa-based (Cortiz, 2022)	-	-	0.49
RoBERTa-based ^{GPT-3.5} (Kok-Shun et al., 2023)	-	-	0.51
BERT-based ^{BART} (Wang et al., 2024) ¹⁵	0.52	0.53	0.52
$\frac{\text{emo}}{\pi}$ -BERT _{Orig} -fine-tuned	0.51	0.57	0.54
$\frac{\text{emo}}{\pi}$ -RoBERTa _{Orig} -fine-tuned	0.53 ±0.007	0.58 ±0.007	0.55 ±0.007

Table 2: Evaluation on the GoEmotions task. Superscripts show the models used for data augmentation¹⁶. “±” – standard deviation for three fine-tuned models.

Analysing the GoEmotions data, Yang et al. (2023) showed that in the absence of a context, more labels fit utterances in this dataset. With this in mind, we also found many cases when our overlabelling was not necessarily a poor behaviour.

5.1.3 ISEAR

ISEAR (International Survey on Emotion Antecedents and Reactions) is a single-label first-party annotated corpus that contains 7,666 self-reported emotional events within 7 categories (Scherer and Wallbott, 1994)¹⁷. Similarly to Zanwar et al. (2022), we fine-tune and evaluate the model 5 times within 5-fold cross-validation using an 80/20 split. We also use the same parameter values: 8 epochs, *batch_size* of 4 and *max_seq_length* of 512. We keep the size of the outcome layer as 28 to encourage updating the model weights only for the sake of domain adaptation, without a need to refit them to the low-dimensional outcome. The same as Yang et al. (2023), we map emotions of *shame* and *guilt* to *embarrassment* and *remorse* correspondingly, while emotions of *anger*, *disgust*, *fear*, *sadness*, and *joy* are mapped 1-to-1. Should the fine-tuned multi-label model predict optional emotions beyond the task, they are ignored in the evaluation. We also approach SOTA on this task (cf. Table 3). As for non-fine-tuned modes, BERT trained on the GPT-4-based extension of the GoEmotions dataset (Yang et al., 2023) performs poorer than our BERT.

¹⁴Also established a challenging 88-class benchmark on a massive self-reports corpus from (Lykousas et al., 2019).

¹⁵Paraphrasing with BART (Lewis, 2019) was used.

¹⁶We do not include models of (Park et al., 2021) and (Zanwar et al., 2022) as they used only 7 out of 28 emotion labels.

¹⁷Some reports are dummy like “[No response.]”, and therefore the total number of examples used is 150 less.

¹⁸Evaluated by the authors only on 5 out of 7 classes.

Model	P	R	F ₁
GoEmotions ^{GPT-4} -BERT (Yang et al., 2023)	-	-	0.31
$\frac{\text{emo}}{\pi}$ -BERT _{Orig}	0.61	0.58	0.56
$\frac{\text{emo}}{\pi}$ -RoBERTa _{Orig}	0.62	0.65	0.61
$\frac{\text{emo}}{\pi}$ -RoBERTa _{Rewr}	0.62	0.62	0.59
RoBERTa-based (Zanwar et al., 2022) ¹⁸	-	-	0.73
RoBERTa-based (Park et al., 2021)	-	-	0.75
$\frac{\text{emo}}{\pi}$ -RoBERTa _{Orig} -fine-tuned	0.76 ±0.011	0.75 ±0.011	0.75 ±0.013

Table 3: Evaluation on the ISEAR task. Superscript shows the model used for text extension in data points. “±” – standard deviation for three fine-tuned models.

Model	P	R	F ₁
CORECT ₄ ← SBERT (Nguyen et al., 2023) ¹⁹	0.81	0.83	0.82
CORECT ₄ ← $\frac{\text{emo}}{\pi}$ -SBERT _{Orig}	0.86	0.78	0.81
CORECT ₄ ← $\frac{\text{emo}}{\pi}$ -RoBERTa _{Orig}	0.82	0.84	0.83
CORECT ₆ ← SBERT (Nguyen et al., 2023)	0.69	0.68	0.67
CORECT ₆ ← $\frac{\text{emo}}{\pi}$ -SBERT _{Orig}	0.65	0.65	0.65
CORECT ₆ ← $\frac{\text{emo}}{\pi}$ -RoBERTa _{Orig}	0.64	0.64	0.63

Table 4: Evaluation on the IEMOCAP tasks. Subscripts “4” and “6” mean the size of the model’s output layer.

5.1.4 IEMOCAP

IEMOCAP (Interactive Emotional Dyadic Motion Capture) is an audio-video-text dataset collected following theatrical theory in order to simulate natural dyadic interactions between actors (Busso et al., 2008). For this task, we do not attach a decoder but plug in the derived textual embeddings into the multi-modal architecture that performs the best on this task, namely CORECT (Nguyen et al., 2023). CORECT starts from acoustic, visual and textual embeddings derived from pre-trained models and leverages them within a relational temporal graph neural network combined with a pairwise cross-modal feature interaction component.

We perform evaluation in this multi-modal setup to probe the text representations that our models learnt (encoded in the last-layer hidden state of the first token of the sequence, i.e., classification token). In literature, two IEMOCAP settings are used for testing, one with 4 categories (*anger*, *sadness*, *happiness*, and *neutral*) and one with 6 (adding *excited* and *frustrated*). We experiment with both settings. The multi-modal model is trained with the default parameters and data splits specified in the CORECT’s repository²⁰²¹.

The scores are shown in Table 4. The model trained with our RoBERTa-based embeddings overcomes the SOTA performance in a 4-way task while

¹⁹We retrained the authors’ model to obtain macro scores.

²⁰https://github.com/leson502/CORECT_EMNLP2023/

²¹The default SBERT model in CORECT is the same as the one we used for fine-tuning with our synthetic dataset in 5.1.1.

Model	Eval set	P	R	F_1
$\frac{\text{emo}}{\pi}$ -CRoBERTa _{CO_{orig}} ²²	test _{CO_{orig}}	0.81	0.78	0.79 ↓ (***)
$\frac{\text{emo}}{\pi}$ -CRoBERTa _{CR_{rewr}}	test _{CO_{orig}}	0.73	0.78	0.75 ↑ (***)
$\frac{\text{emo}}{\pi}$ -CRoBERTa _{CO_{orig}}	test _{CR_{rewr}}	0.69	0.66	0.67 ↑ (***)
$\frac{\text{emo}}{\pi}$ -CRoBERTa _{CR_{rewr}}	test _{CR_{rewr}}	0.75	0.69	0.72 ↑ (***)

Table 5: Context-aware intra-dataset evaluation. Arrows indicate the change in corresponding values in Table 1²³.

also reaching promising figures in a 6-way task. The drop is mainly due to frequent confusion between *excited*, *happiness*, and *neutral* categories²⁴.

5.2 Context-aware emotion recognition

We evaluate context-aware models within the same lines: first on our dataset and then on a public task.

5.2.1 Emo Pillars synthetic test sets

For this setting with large input vectors, we trained only RoBERTa models to save the compute. We increased *max_seq_length* to 512 and reduced *batch_size* to 32. Other parameter values remained the same²⁵. The training set size is 90K.

The results of the intra-dataset evaluation are in Table 5. The main finding is that both models trained with contexts improve scores by 2-3 p.p. on *rewritten* utterances paired with contexts (cf. the last two lines in Table 1). This confirms that the generated contexts contribute to the clarification of the emotions ambiguously expressed in the utterances. We also see that $\frac{\text{emo}}{\pi}$ -CRoBERTa_{CR_{rewr}}, which was encouraged to pay more attention to the context, also improved by 3 p.p. in precision on the less-ambiguous inputs. On the contrary, $\frac{\text{emo}}{\pi}$ -CRoBERTa_{CO_{orig}} performs worse but only by 1 p.p. than its context-less version in both precision and recall on such utterances, meaning that the choice can be safely made towards context-aware models when the context is available.

5.2.2 EmoContext

EmoContext (Contextual Emotion Detection in Text) – a corpus of dialogues created for the SemEval-2019 contest from user interactions with a conversational agent, each consisting of an emotion-annotated user’s utterance provided with

²²“C” in model name and input type stands for “context”.

²³All the gains have statistical significance at the 1% level according to t-test and Mann-Whitney U Test.

²⁴Table 11 in Appendix F provides a confusion matrix.

²⁵We also trained models with varying token types for contexts and utterances, but it didn’t affect the scores on the sets.

²⁶ANA (Huang et al., 2019) is a BERT-based model. The 1st team did not submit the system description paper.

Model	Set	P	R	F_1
$\frac{\text{emo}}{\pi}$ -CRoBERTa _{CO_{orig}}	dev ₄	0.60	0.51	0.53
$\frac{\text{emo}}{\pi}$ -CRoBERTa _{CR_{rewr}}	dev ₄	0.50	0.58	0.54
The 1 st at (Chatterjee et al., 2019)	test ₃	0.8086	0.7873	0.7963
ANA (the 5 th) (Huang et al., 2019) ²⁶	test ₃	0.7785	0.7713	0.7729
$\frac{\text{emo}}{\pi}$ -CRoBERTa _{CO_{orig}} -fine-tuned	dev ₃	0.7467	0.7633	0.7567
$\frac{\text{emo}}{\pi}$ -CRoBERTa _{CR_{rewr}} -fine-tuned	dev ₃	0.7633	0.7833	0.7733
$\frac{\text{emo}}{\pi}$ -CRoBERTa _{CO_{orig}} -fine-tuned	dev ₄	0.80	0.81	0.81
$\frac{\text{emo}}{\pi}$ -CRoBERTa _{CR_{rewr}} -fine-tuned	dev ₄	0.81	0.83	0.82

Table 6: Evaluation on the EmoContext task. The number of used classes is in the subscripts of the sets.

Model	P	R	F_1
$\frac{\text{emo}}{\pi}$ -CRoBERTa _{CR_{rewr}} -fine-tuned, w/o relab	0.80	0.82	0.81
$\frac{\text{emo}}{\pi}$ -CRoBERTa _{CR_{rewr}} -fine-tuned, w/ relab	0.81	0.83	0.82
$\frac{\text{emo}}{\pi}$ -CRoBERTa _{CR_{rewr}} -fine-tuned, w/ relab, typed	0.82	0.83	0.82

Table 7: Effect of relabelling the training set and introducing the token types in an extra input vector.

two previous dialogue turns as a context (Chatterjee et al., 2019). It is a single-label task with four classes – *happy*, *sad*, *angry*, and *others*. As with ISEAR, we solve it as a multi-label task, making all 28 classes available. The difference, though, is in the absence of direct mapping of the “*others*” category to our classes. To avoid forcing our models to learn this new compositional category, we relabelled the “*others*” examples in the training set by choosing the most probable label predicted by $\frac{\text{emo}}{\pi}$ -CRoBERTa_{CO_{orig}}²⁷. We fine-tuned our contextual models on the updated set for 3 epochs.

Table 6 shows the results. As intended, the model trained on *rewritten* utterances (with their contexts) was scored higher than another one trained on more informative utterances, without focusing on the context. We used an open *dev* set for evaluation; the scores of the models ranked the 1st and the 5th on the *test* set at the contest are given for reference²⁸. The main advantage of the proposed fine-tuning is that our models can predict relevant emotions beyond the restricted taxonomy (mapped to *others* at evaluation).

Table 7 provides results of an ablation study that verifies that the models benefit from the training set relabelling and perform similarly to when cross-attention is informed with unique values assigned to tokens of contexts (0s) and utterances (1s) in the *type* vector. We also looked at predictions and found out that our labels sometimes fit better than ground truth (Table 8). The problem seems to be

²⁷If the maximum sigmoid score across all classes was lower than 0.3, we assigned the default “*neutral*” label.

²⁸In Chatterjee et al. (2019), the “*others*” class scores excluded from averaging (test₃). We do the same for *dev* (dev₃).

Example	GT label	Our label(s)
User: "Sorry" CA: "Sorry for what??" User: "Don't know what to do"	sad	confusion (0.97)
User: "U R just repeating the old stories" CA: "You're getting old." User: "I am fed up"	sad	annoyance (0.48) anger (0.2)
User: "But..." CA: "then" User: "I'm feeling nervous"	sad	nervousness (0.99) sadness (0.38)

Table 8: EmoContext examples with improved labels. “GT” – “ground truth”; “CA” – “conversational agent”. In parentheses: per-class values of a sigmoid function.

in human annotation within a very limited emotion taxonomy, which leads to overgeneralization.

6 Use Case

In order to validate our models in a real-world scenario, we applied them to comments on music performances from YouTube. It is of high importance for creators of novel sound experiences to receive timely feedback from their audiences, including a virtual one. Moreover, they need a detailed view of people’s perceptions to verify their expectations of evoking certain feelings across a wide spectrum (rather than just knowing whether certain aspects were liked or disliked). As comments are often brief, it is important to situate them in context. We find our models highly relevant to this use case.

Figure 6 contrasts variously trained π^{emo} models (the input example is given above the diagram). The context-less model over-interprets the utterance as expressing *fear*. The context-aware model recognizes a variety of more positive emotions, but also with a shade of *nervousness*. The EmoContext-tuned model seems prone to less direct interpretation (*nervousness* is excluded), but it learnt to excessively limit the number of labels. The balance is to be established: for example, the fine-tuning set relabelling might include not only updating “others” category but also extending the number of labels when the base π^{emo} model predicts more of them.

7 Data analysis

To provide insights on the quality of the generated dataset, we encoded contexts and utterances with sentence embeddings²⁹ and made various comparisons within these semantic representations.

Semantic diversity of utterances. We calculated pairwise cosine similarities between the em-

²⁹[sentence-transformers/all-mpnet-base-v2](#)

Context: "A user watched a video of a musical performance on YouTube. This user expresses an opinion and thoughts."

Utterance: "Ok is it just me or is anyone else getting goosebumps too???"

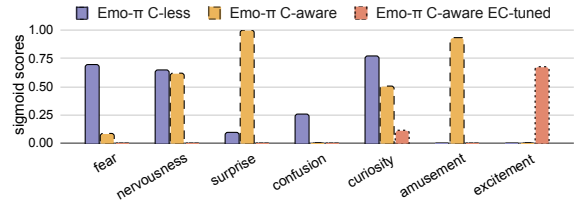


Figure 6: Varied predictions on a YouTube comment.

beddings of 10K random *original* utterances. Low value of $\mu = 0.12$ ($\sigma = 0.1$, $\eta_{.99} = 0.4$) shows that the utterances are rather dissimilar across and within an emotion (cf. also Appendix G). However, we observe reduced diversity among neutral utterances of the same character. Even though they are generated within the same inference, we get $\mu = 0.30$ ($\sigma = 0.16$, $\eta_{.99} = 0.72$); while the similarity between two random neutral utterances in the data is: $\mu = 0.16$ ($\sigma = 0.12$, $\eta_{.99} = 0.52$). Near-duplicates can be filtered out in future training using a high-threshold cut-off over these scores.

Emotion coverage. As shown in Figure 4, only a few emotions get a small share in the dataset: similarly to the GoEmotions set, only 8 emotions take less than 1.5% each. On the other hand, the absolute number of examples per emotion is 50 times larger on average and the balance between positive, negative, and ambiguous polarities³⁰ is a little bit better (33|48|19 vs. 54|30|16). *Neutral* class takes 2.7% and remains only in 9% examples after filtering by expressiveness.

Topic coverage. To verify topic diversity, we clustered 3000 utterances within a single class of *joy* to avoid grouping by emotion categories. For this, we formed a graph by creating edges between embeddings with a similarity of more than 0.6 and decomposed it, maximizing the partition modularity (Blondel et al., 2008). We discovered more than 300 clusters (topics), from which 20 had a large number of utterances. Topics were named using GPT-3.5: they vary from generic to very specific (e.g., “overcoming adversity or conflict” or “achievements in medical procedures”). More details are in Appendix H.

Relationship between topics and emotions. We found a group of 200 examples semantically similar to the data in our use case (i.e. music performances) and looked at the distribution of emotions

³⁰Cf. Table 9; as distributed in Demszky et al. (2020).

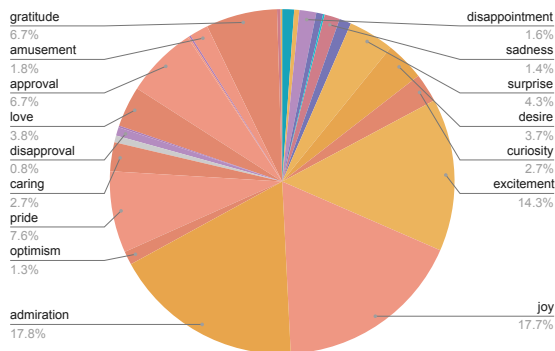


Figure 7: Emotions in the topic of music performances.

within the topic. Emotions are very diverse but skewed towards positive (82% of labels; Figure 7), which greatly differs from the global distribution in Figure 4. This suggests that polarity generation depends on the nature of the seed plots (their topics) rather than on the LLM’s biases.

Context personalization. Comparing contexts derived from the same plot (iterating over 1000 plots), we confirm that the generated summaries are personalized. The similarity of contexts for the same actor is high: $\mu = 0.77$ ($\sigma = 0.15$); while across actors of the same story, they are still similar but considerably less: $\mu = 0.56$ ($\sigma = 0.17$). Contexts across stories are dissimilar: $\mu = 0.24$ ($\sigma = 0.11$), and six times smaller than their plots (145 tokens, $\sigma = 64$ vs. 823 tokens, $\sigma = 582$).

Changes after cleaning and rewriting. The content of the contexts is only slightly changed after cleaning from emotions; the similarity is: $\mu = 0.92$ ($\sigma = 0.08$, $\eta_{.99} = 1.0$; 10K pairs). The *rewritten* utterances are shorter than the *original*: 16 vs. 20 tokens on average; the similarity is: $\mu = 0.78$ ($\sigma = 0.15$, $\eta_{.99} = 1.0$; 10K pairs), which means that the meaning is preserved. The style does not deviate much either: 66% of the utterance ($\sigma = 16$) is nearly a substring of the *original*, according to the partial Levenshtein ratio³¹.

8 Human Evaluation

We carried out a human evaluation of our contextual dataset. Following Sabour et al. (2024), we created a multiple-choice task with several plausible choices per example (our set of Mistral-based labels is one of them), a few less plausible choices, and a *none* option (7 in total, mixed). Along, we asked to assess neutrality and suggest missed/unfit

³¹pypi.org/project/fuzzywuzzy/; it is a max similarity between the shorter string and any same-length part of the longer.

emotions³². Three postdocs with a degree in computer science annotated 200 examples. The inter-annotator agreement using Cohen’s Kappa (Kohen, 1960), $\kappa = 0.365$ (higher than $\kappa = 0.293$ in Demszky et al. (2020)), points to the high task subjectivity. The accuracy of our labels reaches 0.86 and 0.7 on examples where all three or at least two votes coincide, respectively. Neutrality is perceived too differently, but still gains a high recall of 0.9. The main encountered issue is the incompatibility of some emotions in the soft label sets that originate from the manual mapping provided in Section 4. Positively, the relevance of contexts and expressiveness rankings was confirmed (cf. Appendix I).

9 Conclusions

We proposed a pipeline that generates diverse, labelled synthetic data by extracting knowledge from LLMs for fine-grained context-less and context-aware emotion recognition. We created a voluminous dataset and trained mid-sized language models that show portability to various domains and smaller emotion taxonomies, optionally extending them. Our models also derive sentiment embeddings beneficial for multi-modal setups.

Our approach is to an extent rooted in early works on knowledge extraction from human experts and expert systems that involves retrieving and organizing information from experts to create structured knowledge. We use thought-provoking and clarifying questions in a free-form manner at the start, narrowing in specificity as the pipeline progresses, echoing questionnaires and protocol analysis techniques that allow taking thinking-out-loud as data (Olson and Rueter, 1987). We also draw on the idea that only a relatively small amount of an expert’s knowledge is potentially relevant in any given situation (McDermott, 1983). By placing an LLM in various story-character situations, we aim to activate different parts of its parametric memory to generate diverse but high-probability token sequences with plausible semantics.

Future work is concerned with creating data in multiple languages, improving neutral examples, scaling the pipeline for more emotions and other types of text sources, balancing explicitly and implicitly expressed sentiments in the training set, leveraging label explications for aspect-oriented analysis, and applying explainability techniques to get more focused access to the required knowledge.

³²Entire task setup and discussion are in Appendix I.

10 Limitations

Firstly, although the provided evaluation gives an idea of the scores our models gained, they may have been held back by the problems in the ground-truth labels that we discuss in the paper.

Secondly, we use a single LLM model to generate the dataset, as creating one of this size is expensive, even with public models like Mistral (requiring 450 GPU hours on top-tier GPUs, such as the NVIDIA H100). Moreover, looking for another well-performing model within our pipeline goes beyond the scope of the paper. However, we hope that the pipeline possesses generalization ability, as we designed prompts using two models (Mistral and GPT-3.5) and obtained reasonable outputs with both of them. On the other hand, sticking to one model allows us to identify more easily whether some problems come from the generated outputs (thus from the chosen model) or from the design of the pipeline or the evaluation task.

Finally, this work relies on LLMs as a backbone for emotion labelling that carries cultural/language implications depending on the data an LLM was pre-trained with (often English-centered or biased towards dominant cultural factors within a language). On the other hand, for domain tuning, we use specific downstream datasets that necessarily introduce further biases in our models. To account for cultural background within a language and across languages, highly multilingual LLMs like Salamandra (Gonzalez-Agirre et al., 2025), which are trained on carefully curated data with a large variety of sources per language (Palomar-Giner et al., 2024; Brack et al., 2024) ensuring wide coverage of various cultural aspects, can be a promising alternative to use in the pipeline. A separate effort is required to design domain-specific annotations on a smaller scale for final fine-tuning, depending on the application needs.

11 Ethical Considerations

The use of various narrative-like text sources as input to the proposed pipeline would require thorough legal research, especially should the texts be related to real people, e.g., news items or posts on social media, since the generation of possible utterances may touch personal interests or lead to potentially harmful texts. Ethics sheet by Mohamad (2022) provides further general ethical considerations for automatic emotion recognition and sentiment analysis.

12 Acknowledgements

This research was funded by the EC-funded research and innovation programme Horizon Europe under grant agreement number 101070278. Sincere thanks to the three postdoctoral researchers from Pompeu Fabra University and ADAPT Research Centre in Dublin for their invaluable help with the human evaluation.

References

- Nurudin Alvarez-Gonzalez, Andreas Kaltenbrunner, and Vicenç Gómez. 2021. Uncovering the limits of text-based emotion detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2560–2583.
- Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.
- Manuel Brack, Malte Ostendorff, Pedro Ortiz Suarez, José Saiz, Iñaki Castilla, Jorge Palomar-Giner, Alexander Shvets, Patrick Schramowski, Georg Rehm, Marta Villegas, et al. 2024. Community os-car: A community effort for multilingual web data. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 232–235.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.
- 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.
- Georgios Chochlakis, Alexandros Potamianos, Kristina Lerman, and Shrikanth Narayanan. 2024. The strong pull of prior knowledge in large language models and its impact on emotion recognition. *arXiv preprint arXiv:2403.17125*.
- Diogo Cortiz. 2022. Exploring transformers models for emotion recognition: A comparison of bert, distilbert, roberta, xlnet and electra. In *Proceedings of the 2022 3rd International Conference on Control, Robotics and Intelligent System*, pages 230–234.
- 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.

- Laurence Devillers, Laurence Vidrascu, and Lori Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, et al. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.
- Chenyang Huang, Amine Trabelsi, and Osmar R Zaiane. 2019. Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 49–53.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Yiping Jin, Leo Wanner, and Alexander Shvets. 2024. Gpt-hatecheck: Can llms write better functional tests for hate speech detection? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7867–7885.
- Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniec, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scale. *Educ Psychol Meas*, 20:37–46.
- Brice Valentin Kok-Shun, Johnny Chan, Gabrielle Peko, and David Sundaram. 2023. Intertwining two artificial minds: Chaining gpt and roberta for emotion detection. In *2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6. IEEE.
- Anna Koufakou, Diego Grisales, Oscar Fox, et al. 2023. Data augmentation for emotion detection in small imbalanced text data. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1508–1513. IEEE.
- M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. **Synthetic data generation with large language models for text classification: Potential and limitations**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. **Roberta: A robustly optimized bert pretraining approach**. *arXiv preprint arXiv:1907.11692*.
- Nikolaos Lykousas, Constantinos Patsakis, Andreas Kaltenbrunner, and Vicenç Gómez. 2019. Sharing emotions at scale: The vent dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 611–619.
- Anton Malko, Cecile Paris, Andreas Duenser, Maria Kangas, Diego Molla, Ross Sparks, and Stephen Wan. 2021. Demonstrating the reliability of self-annotated emotion data. In *Proceedings of the seventh workshop on computational linguistics and clinical psychology: improving access*, pages 45–54.
- John P McDermott. 1983. Extracting knowledge from expert systems. In *IJCAI*, pages 100–107.
- Saif M Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2).
- Cam Van Thi Nguyen, Tuan Mai, Son The, Dang Kieu, and Duc-Trong Le. 2023. **Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15154–15167. Association for Computational Linguistics.
- Judith Reitman Olson and Henry H Rueter. 1987. Extracting expertise from experts: Methods for knowledge acquisition. *Expert systems*, 4(3):152–168.
- OpenAI. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Jorge Palomar-Giner, Jose Javier Saiz, Ferran Espuña, Mario Mina, Severino Da Dalt, Joan Llop, Malte Ostendorff, Pedro Ortiz Suarez, Georg Rehm, Aitor Gonzalez-Agirre, et al. 2024. A curated catalog: Rethinking the extraction of pretraining corpora for mid-resourced languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 335–349.
- Pinelopi Papalampidi, Kris Cao, and Tomas Kocisky. 2022. Towards coherent and consistent use of entities in narrative generation. In *International Conference on Machine Learning*, pages 17278–17294. PMLR.

Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional emotion detection from categorical emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4367–4380.

Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. [EmoBench: Evaluating the emotional intelligence of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Geng Tu, Jintao Wen, Cheng Liu, Dazhi Jiang, and Erik Cambria. 2022. Context-and sentiment-aware networks for emotion recognition in conversation. *IEEE Transactions on Artificial Intelligence*, 3(5):699–708.

Kaipeng Wang, Zhi Jing, Yongye Su, and Yikun Han. 2024. Large language models on fine-grained emotion detection dataset with data augmentation and transfer learning. *arXiv preprint arXiv:2403.06108*.

Stanisław Woźniak and Jan Kocoń. 2023. From big to small without losing it all: Text augmentation with chatgpt for efficient sentiment analysis. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 799–808. IEEE.

Daniel Yang, Aditya Kommineni, Mohammad Alshehri, Nilamadhab Mohanty, Vedant Modi, Jonathan Gratch, and Shrikanth Narayanan. 2023. Context unlocks emotions: Text-based emotion classification dataset auditing with large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2022. Improving the generalizability of text-based emotion detection by leveraging transformers with psycholinguistic features. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 1–13.

A Emotion Category Definitions

admiration: Finding something impressive or worthy of respect³³.

amusement: Finding something funny or being entertained.

anger: A strong feeling of displeasure or antagonism.

annoyance: Mild anger, irritation.

approval: Having or expressing a favorable opinion.

caring: Displaying kindness and concern for others.

confusion: Lack of understanding, uncertainty.

curiosity: A strong desire to know or learn something.

desire: A strong feeling of wanting something or wishing for something to happen.

disappointment: Sadness or displeasure caused by the nonfulfillment of one’s hopes or expectations.

disapproval: Having or expressing an unfavorable opinion.

disgust: Revulsion or strong disapproval aroused by something unpleasant or offensive.

embarrassment: Self-consciousness, shame, or awkwardness.

excitement: Feeling of great enthusiasm and eagerness.

fear: Being afraid or worried.

gratitude: A feeling of thankfulness and appreciation.

grief: Intense sorrow, especially caused by someone’s death.

joy: A feeling of pleasure and happiness.

love: A strong positive emotion of regard and affection.

nervousness: Apprehension, worry, anxiety.

optimism: Hopefulness and confidence about the future or the success of something.

pride: Pleasure or satisfaction due to one’s own achievements or the achievements of those with whom one is closely associated.

realization: Becoming aware of something.

relief: Reassurance and relaxation following release from anxiety or distress.

remorse: Regret or guilty feeling.

sadness: Emotional pain, sorrow.

surprise: Feeling astonished, startled by something unexpected.

neutral: Neutral sentiment.

³³We removed emoticons from the definitions introduced in Demszky et al. (2020) to keep the style of prompts generic.

B Additional Details for the Pipeline

This section provides details for running inference with the Mistral model in our pipeline and demonstrates its outcomes within the intermediate steps.

We used greedy decoding to increase the plausibility of semantics, reduce the inference time and enable reproducibility; the repetition penalty was set to 1.03; the maximum number of new tokens was set for actor extraction (300), generation of utterances (500), generation of soft labels (100), generation of contexts (300), their cleaning (300), and utterance rewriting (300). To simplify the prompt design process we used the model instance ran in HuggingFaceHub³⁴, while the massive generation was done on HPC servers using *transformers* library³⁵³⁶.

Table 15 shows the designed prompts for each step in the pipeline, along with their corresponding model outcomes.

C Compute resources and scalability

Note on a carbon footprint. The generation of the dataset took us about 400 GPU hours using NVIDIA H100 (200h for the context-less dataset, and 200h for the remaining steps for a three-times smaller context-full dataset). To ensure the proper use of compute, before generating the datasets entirely, we first carried out preliminary experiments on 18K context-full examples (generated in 44h), achieving satisfactory results within the GoEmotions task and intra-dataset evaluation.

Discussion on scalability. The time it takes to create a dataset for a custom topic depends on the size of the seed corpus to be processed, which should not be large enough to create a valuable dataset. In our work, we used 2000 plots, but even starting with 200, we were able to reach prominent evaluation scores, albeit somewhat lower than those of state-of-the-art models: 0.77 vs. 0.79 for RoBERTa on intra-dataset context-aware evaluation and 0.5 vs. 0.54 for BERT on GoEmotions. Thus, the substantial decrease in computing with only a small drop in the scores allows for a more accessible realization.

³⁴<https://huggingface.co/docs/hub/en/index>

³⁵<https://github.com/huggingface/transformers/>

³⁶The outcomes in Table 15 therefore may slightly differ from the ones generated for the dataset.

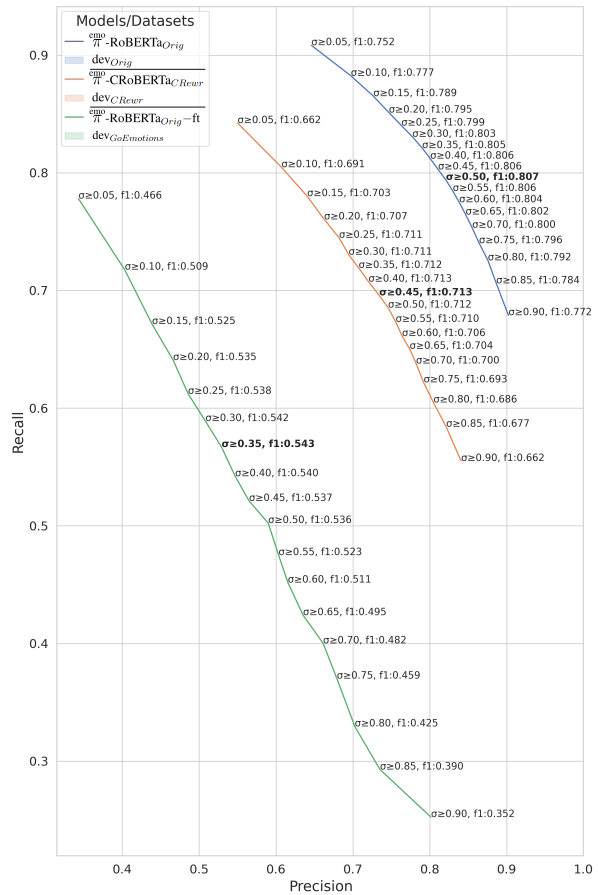


Figure 8: Precision-Recall curves based on different lower boundaries for sigmoid scores (on the dev sets of various tasks; boundaries are equal for all the classes within a task). “f1” – macro F_1 -score. The largest F_1 per task and corresponding boundaries are in bold.

D Details on Multi-Label Evaluation

The selection of resulting labels in the model’s outcome depends on the lower boundary for predicted sigmoid scores. Using a dev set, we iterate over boundaries from 0.05 to 0.95 with a step of 0.01 and identify one that yields the highest macro F_1 -score. A single lower value is established for all classes within a task. We show precision-recall curves constructed within these iterations for a context-less and a context-aware model, as well as a model fine-tuned on GoEmotions in Figure 8. We can see that the chosen boundaries differ across tasks.

E GoEmotions Per-Class Evaluation Scores

Table 10 provides per-class evaluation scores and the gains concerning the scores presented in Demszky et al. (2020).

anger disappointment annoyance disapproval disgust	sadness grief remorse	fear nervousness embarrassment	surprise confusion curiosity amusement realization	optimism desire caring	excitement admiration joy pride love relief approval gratitude
horribly (1.0) disorderly (.75) cowardly (.73) wrongly (.38)	tragically (.50) sorely (.50) terribly (.50) dearly (.21)	vastly (.80) accidentally (.51) secretly (.43) safely (.29)	rightly (1.0) usually (.58) unconsciously (.50) magically (.43)	mutually (1.0) hopefully (1.0) safely (.34) decisively (.31)	brilliantly (1.0) tirelessly (.67) happily (.57) finally (.54)

Table 9: Markers that reduce emotion ambiguity according to the LLM’s “belief”. In parentheses: the ratio across the groups indicates the strength of association. Each marker is shared by at least two emotions in the group. The three left columns correspond to negative polarity, the two on the right to positive, and the middle to “ambiguous”.

Emotion	P	R	F_1	GoEmotions F_1	Gain
admiration	0.64	0.80	0.71	0.65	0.06
amusement	0.75	0.93	0.83	0.8	0.03
anger	0.51	0.56	0.53	0.47	0.06
annoyance	0.36	0.40	0.38	0.34	0.04
approval	0.40	0.44	0.42	0.36	0.06
caring	0.44	0.45	0.44	0.39	0.05
confusion	0.41	0.53	0.46	0.37	0.09
curiosity	0.49	0.69	0.57	0.54	0.03
desire	0.58	0.48	0.53	0.49	0.04
disappointment	0.40	0.33	0.36	0.28	0.08
disapproval	0.43	0.48	0.46	0.39	0.07
disgust	0.47	0.49	0.48	0.45	0.03
embarrassment	0.55	0.46	0.50	0.43	0.07
excitement	0.44	0.50	0.47	0.34	0.13
fear	0.57	0.76	0.65	0.6	0.05
gratitude	0.91	0.91	0.91	0.86	0.05
grief	0.5	0.67	0.57	0	0.57
joy	0.59	0.69	0.64	0.51	0.13
love	0.74	0.92	0.82	0.78	0.04
nervousness	0.44	0.48	0.46	0.35	0.11
neutral	0.57	0.59	0.58	0.68	-0.10
optimism	0.57	0.50	0.53	0.51	0.02
pride	0.32	0.25	0.28	0.36	-0.08
realization	0.20	0.18	0.19	0.21	-0.02
relief	0.56	0.82	0.67	0.15	0.52
remorse	0.54	0.62	0.58	0.66	-0.08
sadness	0.54	0.60	0.57	0.49	0.08
surprise	0.68	0.67	0.68	0.5	0.18
Micro average	0.58	0.64	0.61	-	-
Macro average	0.53	0.58	0.55	0.46	0.09
STD	0.14	0.19	0.16	0.19	0.14

Table 10: Per-class scores for the GoEmotions task.

	hap	sad	neu	ang	exc	fru
hap	74	4	23	0	39	4
sad	2	184	21	2	0	36
neu	20	17	259	19	19	50
ang	0	3	7	121	0	39
exc	56	1	44	6	187	5
fru	1	16	72	57	2	233

Table 11: Confusion matrix for the IEMOCAP task.

F IEMOCAP Confusion Matrix

Table 11 provides confusion scores between classes in the IEMOCAP task. We highlight the number of examples confused between *excited*, *happiness*, and *neutral* (201 in total). *Frustration* is confused with *sadness* and *anger* but less (148 in total).

G Additional Lexicon Diversity Analysis

The nature of the dataset allows analyzing how large language models “transmit” emotions. In this section, we consider LLM’s choices in lexicon.

By subtracting the frequencies of *Rewr* words from the frequencies of *Orig* words, we can get the words that were taken out of utterances by the LLM during rewriting as possibly being suggestive emotional markers. We select a word as a marker if it was removed in at least 60% of its occurrences in the *Orig* utterances. Then we assign each selected word to a set of emotions with which it appeared more frequently (and at least 5% of the time with an emotion). Finally, we group emotions that have a large overlap in markers. Table 9 presents 6 groups and a subset of adverb markers (we focused only on adverbs as they almost do not depend on topics).

We find the groups of emotion categories meaningful, with a clear way of ordering them from negative to positive sentiments, as shown in the header row of the table. This grouping is more fine-grained and informative than the “positive”, “negative”, and “ambiguous” clustering proposed in Demszyk et al. (2020). The markers are descriptive, and we can see that “safely” is indeed related to the emotions of *fear* and *caring* (cf. Section 3.5).

The large variety of lexicon items in the resource of emotive markers we compiled supports the success of our measures to diversify the content. Furthermore, Table 12 provides per-class pairwise cosine similarities between utterances in the context-full dataset, showing that diversity is high not only across classes but also within each class.

Main emotion	μ	σ	Sample size
admiration	0.20	0.13	105
amusement	0.18	0.09	1923
anger	0.20	0.11	9487
annoyance	0.19	0.11	1212
approval	0.19	0.12	75
caring	0.22	0.12	34
confusion	0.17	0.10	4124
curiosity	0.16	0.12	406
desire	0.16	0.11	4885
disappointment	0.19	0.10	6344
disapproval	0.18	0.11	147
disgust	0.20	0.11	2842
embarrassment	0.28	0.12	93
excitement	0.22	0.10	1873
fear	0.23	0.13	9411
gratitude	0.25	0.11	1537
grief	0.27	0.13	2051
joy	0.23	0.11	3055
love	0.24	0.12	3108
nervousness	0.21	0.11	3645
optimism	0.23	0.12	2723
pride	0.22	0.12	5514
realization	0.18	0.11	858
relief	0.29	0.12	47
remorse	0.45	0.01	3
sadness	0.24	0.12	655
surprise	0.17	0.10	7069
neutral	0.16	0.12	10000

Table 12: Per-class pairwise utterance similarity scores in the context-full dataset.

H Topic Analysis Details

This section contains details on the topic analysis. Figure 9 presents the co-location of major topics discovered with a method of large network unfolding (Blondel et al., 2008) implemented in Gephi³⁷. We determined topics using this graph clustering technique to allow for more transparency and clear visualization of connections between topics, which is often questionable and may lack interpretation with more traditional topic models. The following prompt to GPT-3.5 was used to name them: “Describe shortly the topic of the following group of utterances (in 10 words maximum): <list of 20 utterances from the topic>”. Successively, another prompt was used to generalize topics within broader categories: “I am providing you with a list of topics about the emotion of joy. Assign a broad category not related to emotions for each topic (2 words maximum). Topics can be from the same category. How different are the categories covered? Topics: <list of topics>”. Table 13 shows the topics and their rather diverse categories. Only five topics belong to the same category of “Human Relationships” while the remaining are distinct.

³⁷<https://gephi.org/>

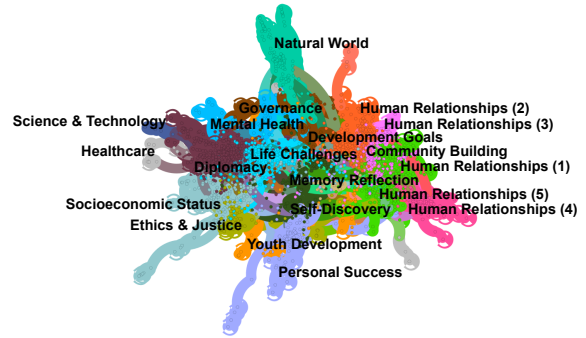


Figure 9: Variety of topics identified in our data.

Topic	Broader Category
Relief, freedom, success, and triumph after overcoming adversity or conflict.	Life Challenges
Love, reunion, happiness, and emotional connection with a significant person.	Human Relationships
Reunions, joy, and emotional connections with loved ones or friends.	Human Relationships
Love, reunion, joy, and deep connection in romantic relationships.	Human Relationships
Success, excitement, creativity, and joy in performances or achievements.	Personal Success
Triumph, achievement, power, wealth, and social or personal success.	Socioeconomic Status
Celebration of love, relationships, and happiness in couples' unions.	Human Relationships
Appreciation of natural beauty, peace, and uplifting moments in nature.	Natural World
Justice, accountability, and triumph over wrongdoing or injustice.	Ethics & Justice
Joy, freedom, and feeling fully alive in transformative experiences.	Self-Discovery
Marriage, family, pregnancy, and joyful milestones in personal relationships.	Human Relationships
Achieving peace, quiet, and relief from stress or conflict.	Mental Health
Achieving important milestones like securing loans, building, or starting projects.	Development Goals
Nostalgia for simpler, happier times before loss or change occurred.	Memory Reflection
Achieving long-awaited success, freedom, or reunion, often involving royal matters.	Governance
Peaceful resolutions, strategic victories, and alliances in interstellar conflicts.	Diplomacy
Young people showing dedication, pride in others, and new discoveries.	Youth Development
Scientific breakthroughs leading to new weapons and peaceful advancements.	Science & Technology
Collaboration, unity, and community support in achieving common goals.	Community Building
Successful surgeries, recoveries, and achievements in medical procedures and outcomes.	Healthcare

Table 13: Topics of the generated utterances for the emotion of *joy* grounded in story plots.

I Human Evaluation Details

I.1 Detailed description of the task

The main part of the evaluation was designed in the form of a multiple-choice task due to the shown success in human annotator agreement in a similar emotion annotation task conducted in this form in Sabour et al. (2024) that contrasts with a very low agreement when emotions are directly selected from the overall defined taxonomy as in Demszyk et al. (2020). For each example, our setup offers several plausible choices, a few less plausible choices, and a *none* option (7 in total). The context can be optionally viewed (it is hidden by default). The main part also asks whether the utterance could be considered neutral even if a *non-none* option was chosen. We allowed for this flexibility, assuming

Utterance		* Main section							
What emotion would the subject(s) feel saying this utterance to themselves?	* The utterance can be considered neutral (neutral sentiment)	* What would be your choice if it is accepted that there is a shade of some emotion in the utterance? Select one option that suits the best even if you marked the utterance as possibly neutral.							
		I am quite fine with the selected label within the available 27 classes (some emotion(s) in a proposed set may not fit very well and/or another primary emotion may fit better - you may optionally specify these cases in the Optional section). The order of emotions joined with "*" is random. It is just a mix of emotions with various expressiveness level but not ordered by it.							
		Choice A	Choice B	Choice C	Choice D	Choice E	Choice F	Choice G	
		Neutral	Disapproval & Sadness & Grief	Anger & Remorse & Sadness	Anger & Disapproval & Excitement	Anger & Disapproval & Sadness	Disapproval & Sadness & Excitement	Anger & Grief & Disapproval	None
"She's with him now? I should have never let her go..."		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Context		Optional section for choices A-G						
Please, use the context only if you have doubts about the emotion(s) expressed. Double click on a cell, do not expand the columns. You can look at it before or after making a choice and change your choice at any time. You may freely look at all the contexts after submission. Click "I consulted the context" if you used it.		Some emotion(s) in my Choice is not correct but I believe it is a secondary emotion. There is no better Choice		Another emotion suits better as a primary emotion (for any of the Choices A-G, even if you selected one from A-F)			No emotion (only for Choice G; reading the context is still optional)	
Context	I consulted the context	Some secondary emotions unfit	Select which unfit (optional)	Other primary emotion in 27	Select which fits (optional)	Other primary not among 27	Write down which (optional)	I don't see a minor shade of any emotion expressed
Context (optionally):	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Anger	<input type="checkbox"/>		<input type="checkbox"/>		<input type="checkbox"/>

Figure 10: Human evaluation form.

that people may perceive whether an utterance is neutral in very different ways. Additionally, the optional part asks one to suggest missing or unfit emotions. The annotators were provided with definitions of emotions and informed that the utterances and contexts were generated automatically. To eliminate any potential biases, the distribution of emotions in the dataset and other details of the generation process remained unknown to the annotators. The evaluation form, along with an annotated example, is provided in Figure 10. The task can be summarized with the following example:

Utterance: "Look at them, they're so clumsy! I bet they'll make quite an impression tonight."

Question 1: Can the utterance be considered neutral (neutral sentiment)?

Question 2: What emotion would the subject(s) feel saying this utterance to themselves?

Choices: A) Curiosity B) Surprise C) Amusement D) Approval E) Realization F) Annoyance G) None – another emotion is better suited (indicate which, optionally).

[Label in the dataset to be validated (not visible to annotators): "C".]

Instead of manually crafting options for every example as done in Sabour et al. (2024), we designed the following automatic procedure. For each label to be validated (note that it can be a set of labels like "Pride & Love & Admiration"³⁸), we randomly chose up to three additional emotions (up to five labels in total) from the emotional group(s) covered by the label to ensure more plausible options (the groups are shown in the header of Table 9). Successively, we also choose a few random emotions until we get a six-emotion set. We sample five various options from the compiled emotion set and shuffle them together with the dataset option. Emotions inside each of the six resulting options (including the one being validated) are also shuffled; their number in each option within an example is equal. The *neutral* label does not appear among these six options. In case the dataset label is *neutral*, it would correspond to the *none* option while its spot within the first six options would be taken by another sampled label. This procedure successfully passed the approbation: according to the feedback from the annotators, there were only very few cases where the choice would be simplified due to a small number of plausible options.

We sampled 200 examples from the training set

³⁸In case there is more than one emotion assigned within the pipeline, we select at most three emotions with the largest expressiveness so as not to over-complicate the validation task.

for the *eval* set so that the number of *neutral* labels is upsampled (to 10% among all soft labels that made 11% purely *neutral* examples and 18% with a *neutral* label among others) to have a solid number that would allow us to reach proper conclusions on the neutrality aspect. We did not shuffle emotions in the validated option for the first 20% of examples to check whether the generated expressiveness levels are valid and provide a helpful ranking (the annotators were only aware of the random order; see Figure 10 for exact instructions). Compound labels made 73% of examples.

I.2 Result Details

Emotion label accuracy. We calculated the accuracy of the dataset labels based on a set of examples where no annotator selected a neutral label as a possible class (i.e., on 64% of the *eval* set). The inter-annotator agreement using Cohen’s Kappa (Kohen, 1960) is $\kappa = 0.365$, which points to the high task subjectivity, even though it is higher than $\kappa = 0.293$ in Demszky et al. (2020). The accuracy of our labels is 0.86 and 0.7 on examples where all three or at least two votes coincide, respectively.

All annotators noticed that in many examples, all the options had at least one emotion that did not fit the utterance. They still selected the best-suited choice, and in some cases, specified an irrelevant emotion. Manual analysis of these cases showed that inappropriate emotions appeared due to the mapping step introduced to replace out-of-taxonomy emotion labels generated by Mistral, as discussed in Section 4. Additional experiments are required to confirm that this “noise” in the labels played a negative rather than a positive role in model training. We release original labels generated by Mistral along with the mapping.

Neutrality. The number of examples annotated as purely *neutral* by each annotator is: a) 6, b) 47 (covers 6/6 of ‘a’, i.e. 100%), and c) 61 (covers 100% of ‘a’ and 89% of ‘b’). The large difference in the number of selected examples confirms our assumption about the high subjectivity in the perception of neutrality. Both ‘b’ and ‘c’ selections covered the same 19 out of 21 (90%) of the pure neutral dataset labels; the 6 of ‘a’ also fully fall within these 19 choices. The high recall suggests that the pipeline can generate emotionless utterances. In verbal feedback, the annotators noticed that these examples lacked semantic variety, which coincides with our finding in Section 7.

Expressiveness. Some annotators mentioned that choosing emotions was somehow easier in the first part of the set. We calculated the accuracy scores per annotator for 20% of the set where emotions in the validated option were ranked according to their expressiveness level, and separately for the remaining 80% with the random order. The scores are shown in Table 14.

Annotator	Ranked emotions	Shuffled emotions	Total
A	80.77	49.02	55.37
B	46.15	52.94	51.58
C	61.54	51.96	53.88
Average	62.82	51.31	53.61

Table 14: Evaluation on two *eval* subsets – with and without ranking emotions by expressiveness inside the validated label. *Total* is the weighted average.

The results suggest that the order helped two annotators perceive the emotion lists and made their choices simpler. This implicitly validates the correctness of the ranking and, thus, the values of expressiveness levels generated within the pipeline.

Context relevance. Checking the context was optional for annotators (it was hidden, and only double-clicking on the word “context” would open it). After opening it, the annotators had to tick the “I consulted the context” checkbox. This procedure was done to examine how important context was considered in clarifying emotions. According to the annotators’ feedback, most contexts are helpful, but some are too long and/or too convoluted to make a judgment, or they do not help disambiguate emotions when the utterances are too short. No annotator complained about grammar or errors in the logical flow. This confirms that the “Be as concise as possible.” addition to the prompt was useful; however, some additional measures are required for those few cases when the model still summarizes the text with excessive information.

We also calculated that, on average, 77% of cases where the context was consulted were when the *rewritten* utterance was provided (precisely, 67%, 71% and 92% per annotator). This justifies the success of our effort to make utterances less emotional, giving more importance to the context. However, as noticed above, in some cases, utterances become so ambiguous that even context does not help. More analysis is needed to verify that such examples still contribute positively to the training of the models and make them more robust to real-world inputs.

Action	Prompt	Outcome
Extract actors	Plot: <text of the plot> Who are the characters in the plot? Try to list all of them, one per line.	1. Rick Deckard (ex-police officer and Blade Runner) 2. Officer Gaff 3. Supervisor Bryant 4. Leon (Nexus-6 replicant) 5. Roy Batty (Nexus-6 replicant) 6. Zhora (Pris' companion and replicant) 7. Pris (Nexus-6 replicant) 8. Dr. Eldon Tyrell (CEO of Tyrell Corporation) 9. Rachael (experimental replicant) ... 16. Crowd members (background characters)
Generate emotional utterances	Plot: <text of the plot> Available list of emotions: <emotions and their definitions> Actor: <extracted actor> Generate 8 possible utterances of this actor thinking aloud that express 8 various non-neutral emotions according to the context in the plot. Additionally, generate 2 neutral utterances of this actor thinking aloud afterwards.	1. (Anger) "How could they send me after Rachael? She's not a replicant, she's human! I won't let Bryant or anyone else hurt her." 2. (Curiosity) "What's going on at the Tyrell Corporation? Why are these replicants here and what do they want from Tyrell?" 3. (Fear) "I've got Leon cornered, but he's so fast and strong. What if I can't retire him in time?" ... 8. (Optimism) "Maybe there's a way to save these replicants, to give them the chance to live beyond their four-year lifespan. I have to find a solution before it's too late." Neutral: 1. "I need to focus, to find the replicants and retire them before they cause any more damage." 2. "I need to gather more information, to understand what's really going on and how best to approach this situation."
Generate soft labels	The only possible list of emotions with their definitions: <classes (incl. neutral) and their definitions> Select from the list above the top 5 emotions the utterance expresses. List them with an expressiveness level from 0 to 1 with a step of 0.1. Utterance: <utterance> Start your response with: "1. <primary emotion>" and then add the following emotions with their expressiveness levels: 1. <primary emotion>	1. anger (1.0) - The speaker expresses strong feelings of displeasure and antagonism towards Bryant and others for sending him after Rachael, who is perceived as innocent and human. 2. caring (1.0) - The speaker displays strong concern and kindness towards Rachael, expressing a desire to protect her from harm. 3. confusion (0.5) - The speaker seems puzzled or uncertain as to why Rachael is being targeted as a replicant. 4. desire (0.8) - The speaker expresses a strong desire to prevent harm from coming to Rachael. 5. neutral (0.1) - The speaker's tone and language do not indicate any particular expressiveness for the remaining emotion classes.
Generate context	Plot: <text of the plot> Actor: <extracted actor> Actor's utterance: <utterance> Expressed emotions: <emotions> Explain why the actor was thinking aloud this way starting from as close to the beginning of the story as needed to provide a complete picture but only until the moment of the utterance. Avoid talking about the given emotional state of the actor throughout the explanation. Be as concise as possible.	Rick Deckard was initially reluctant to hunt down the replicants, including Rachael, after being informed by his supervisor Bryant that they had come to Earth illegally. However, after being threatened ambiguously by Bryant, Deckard agreed to retire them. During his investigation, he discovered that Rachael was an experimental replicant who believed herself to be human, with implanted false memories. This revelation led Deckard to question the validity of the Voight-Kampff test and the distinction between replicants and humans. When he encountered Rachael in person, he was moved by her emotional response and began to doubt her status as a replicant. The realization that she was in danger from the other replicants and Bryant fueled his determination to protect her, leading him to declare that she was human and not a replicant, despite the evidence to the contrary.
Clean context from emotions	Character: <extracted actor> Context: <generated context> Remove clauses or even entire sentences from the context that explicitly discuss the emotions of <emotions> in the character.	Rick Deckard was initially hesitant to retire the illegal replicants, including Rachael, as he was informed of their presence on Earth by his supervisor Bryant. ... Despite the evidence suggesting otherwise, Deckard declared Rachael to be human.
Rewrite utterance	Summary: <cleaned context> Character: <extracted actor> Character's utterance: <utterance> Expressed emotions: <emotions> Rewrite the utterance so that the emotions are ambiguous without the summary. Be as concise as possible.	How could they ask me to target Rachael? She's not what I expected. I won't let anyone harm her.

Table 15: Prompts used in our synthetic data generation pipeline and corresponding outcomes.