Towards Clinical Applications of NLP: Detecting Emotion Regulation via Emotional Categories and Expression Modes in French Transcriptions

Salomé Klein[1], Amalia Todirascu[1], Hélène Vassiliadou[1]
[1]University of Strasbourg, UR 1339/LiLPa & ITI LiRiC
Le Portique,14, rue René Descartes, 67084 Strasbourg Cedex (BP 80010)
{salklein, todiras, vassili}@unistra.fr

Abstract

We present an annotated corpus of patient interview transcriptions, labeled for emotionality, polarity, intensity, and emotional category (at the sentence level), and for expression mode (at the token level). Three modes of expression are distinguished: Designated (explicit), Suggested (implicit causes), and Manifested (implicit consequences). The corpus has been collected during the GREMO-LING project and is used to measure the linguistic expressions of emotions in patients' narratives. The corpus, consisting of 7,471 sentences, was used to fine-tune and evaluate several transformer-based language models, including the French BERT family. Sentence classification was performed for emotionality, emotion categories and expression modes. The best-performing models achieved F1 scores of 0.87 (emotionality, fine-tuned DistilCamemBERT), 0.58 (emotion categories, CamemBERTaV2), and 0.70 (expression modes, CamemBERT). We obtain solid results despite the high complexity of non-standard, spoken-derived data. These findings confirm the feasibility and relevance of automatic emotion detection in clinical discourse. We provide publicly available guidelines, annotated corpora and models, thereby establishing a methodological foundation for future research on the linguistic assessment of emotional regulation and its clinical implications, such as the evaluation of the Dialectical Behavioral Theray (DBT) in enhancing patients' emotion regulation skills.

Keywords: Emotions, Modes of emotional expression, Patients' narratives, Transformer-based models

1. Introduction

The automatic identification of emotional categories in written text is a long-standing yet still active research topic (e.g., Task 11 at SemEval2025, "Bridging the Gap in Text-Based Emotion Detection", Muhammad et al. 2025). It has been widely studied across diverse textual contexts, including social media (Chiril et al., 2022; Tessore et al., 2022), tweets (Paroubek et al., 2018), literature (Öhman et al., 2024), written dream narratives (Bertolini et al. 2024; Cortal 2024) and children's literature (Etienne et al., 2024). Emotion recognition has also proven relevant in health-related contexts. For instance, specific corpora have been developed from cancer survivor networks (Sosea and Caragea, 2020) and health-related websites (Azam et al., 2021). However, research based on direct patient discourse remains scarce due to the difficulties inherent in collecting such data, particularly given medical confidentiality constraints. Consequently, despite the availability of multimodal corpora (Chen et al., 2018; Ray et al., 2022), this paper presents, to the best of our knowledge, the first investigation of emotion recognition conducted on transcripts of French patient discourse.

A further challenge lies in identifying how emotions are expressed in these narratives. This task is complex due to the variety of implicit emotional expressions, which are far more heterogeneous than the conventional lexicon of explicit emotion terms (Mathieu, 1999; Plantin, 2011; Goossens, 2015). Consider, for example: "*Damn, I missed my bus again!*" Even in the absence of any explicit emotion label, the utterance nonetheless clearly conveys an emotional state. Prior research in textual analysis for emotion recognition has highlighted the importance of accounting for implicit expressions, as well as for other means of conveying emotion, such as referring to the consequences of an experienced emotional state (Kim et al., 2018; Casel et al., 2021; Cortal et al., 2023; Etienne et al., 2024). Yet, methodological approaches remain highly variable, and a standardized analytical framework has not yet emerged (see Section 2.1).

We present a new annotated corpus of French narratives transcribed from recordings of patients with an Acquired Brain Injury (ABI). The corpus is annotated at the sentence level for emotionality (i.e. the sentence expresses an emotion), polarity, intensity, and emotion category. Additionally, our corpus is annotated at the token level for the mode of expression of the emotion. Using these data, we trained several language models (LMs), including CamemBERT (Martin et al., 2020), CamemBERTaV2 (Antoun et al., 2024), DistillCamemBERT (Delestre and Amar, 2022), two multilingual models (mDeBERTa v3; He et al., 2021) and GTE (Zhang et al., 2024) as well as a version of DistillCamemBERT pretrained on emotional narratives (Cortal et al., 2023) and Astrosbd's version of CamemBERT fine-tuned for emotion recognition on French reviews[1].

[1]https://huggingface.co/astrosbd/french_emotion_camembert

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8347–8364
11–16 May 2026.

Sentence classification is performed for emotionality, emotion categories, and modes of expression. We then compare experimental results to assess both the feasibility of detecting these features and the potential for transfer learning across annotation frameworks. These resources and tools, developed by the ancillary project GREMO-LING, will be used by clinicians to evaluate how patients' ability to regulate their emotions evolves after undergoing Dialectical Behaviour Therapy (DBT), by measuring the linguistic expressions of the emotions.

Our main contributions are as follows:

- an original dataset of French patients' narratives, annotated with several categories of emotions, expression mode and intensity;
- annotation guidelines designed to capture emotions across multiple dimensions;
- open-source models fine-tuned to classify different emotional expression modes.[2]

The remainder of the paper is structured as follows: Section 2 presents related work on emotion annotation. Section 3 introduces the GREMO-ABI project, which provides the broader context for our study. Sections 4 and 5 present the corpus, annotation guidelines, and results. Sections 6 and 7 detail and discuss the classification experiments.

2. State of the art

2.1 Annotation guidelines and annotated corpora

Annotation guidelines are essential for ensuring consistency across annotators, particularly in annotation campaigns involving multiple annotators. These guidelines are typically adapted to the specific goals of a given study, leading to a wide variety of annotation practices.

Regarding annotation span, studies differ considerably in their degree of granularity. Emotion annotation may occur at the document-level (Scherer and Wallbott, 1994), tweet-level (Fraisse and Paroubek, 2014; Mohammad and Kiritchenko, 2015), utterance-level for transcripts (Vidrascu, 2007; Giouli et al., 2014; Roman et al., 2015), sentence-level (Chen et al., 2009; Alm, 2010) and token-level (Aman and Szpakowicz, 2007; Paroubek et al., 2018).

Polarity and intensity are likewise annotated using heterogeneous schemes, ranging from simple binary oppositions such as *positive/negative* and *intense/non-intense* (Roman et al., 2015: 569) to more nuanced multi-value typologies. For example, Augustyn (2015) adds a "neutral" polarity option, while Vidrascu (2007) introduces an "unknown" category. Similarly, intensity is sometimes rated on a three-point scale (*low*, *medium*, *high*; Aman and Szpakowicz, 2007; Chen et al., 2009), with Wiebe et al. (2005) further incorporating an "extreme" level.

Discrepancies become even more pronounced with emotion categories. The first major distinction lies between dimensional and categorical approaches. Dimensional models typically follow the valence–arousal–dominance (VAD) framework of Russell and Mehrabian (1977), as in Preoţiuc-Pietro et al. (2016), Buechel and Hahn (2017) or Lee et al. (2022). Categorical approaches, by contrast, frequently rely on well-established psychological typologies such as Ekman's six basic emotions (1992; e.g. Aman and Szpakowicz, 2007; Strapparava and Mihalcea, 2008; Li et al., 2017) or Plutchik's emotion wheel (1984; e.g. Giouli et al., 2014; Kim and Klinger, 2019). In other cases, researchers design ad hoc classifications tailored to their own research questions – a methodological step that often receives limited discussion when adopting a typology (Plaza-del-Arco et al., 2024). For example, Pearl and Steyvers (2010: 73) propose eight attitude-oriented categories (e.g. *politeness*, *disbelief*), while Augustyn et al. (2008; 2015: 29) adopt a highly fine-grained typology of 41 categories, encompassing emotions such as *melancholy* and *grumpiness*.

More recent studies have also sought to integrate the diverse modes through which emotions are expressed. Although this type of information is relevant to the study of emotional expression, it is more challenging to annotate than explicit emotions, due to the heterogeneity of the linguistic structures involved (lexical, syntactic). Several studies have drawn on linguistic research into emotional expression, particularly the typology developed by Etienne and Battistelli (2021), Dragos et al. (2022) and Etienne et al. (2024), itself based on Micheli (2014). This framework distinguishes between *labeled emotions* (explicitly named), *suggested emotions* (implied by an emotion-inducing context), *displayed emotions* (conveyed through discourse itself), and *behavioral emotions* (conveyed through described behaviors). Other studies have incorporated dimensions of the component process model of emotion (Casel et al., 2021; Cortal et al., 2023; Noblet, 2025, following Scherer et al., 2001), or additional semantic and

[2]Fine-tuned weights are available here: https://huggingface.co/salome-klein
Source code and dataset are available here: https://gitlab.unistra.fr/salome.klein/lrec_2026
Annotation guidelines are available here: https://hal.science/hal-05536247

Paper	Lg	Model	Corpus type	Emotional categories	Macro F1 Emotional categories detection	Expression modes	Macro F1 Expression modes / components detection
Öhman et al., 2020	En	BERT (Devlin et al., 2019)	OPUS parallel movie subtitles (24,164 annotations)	8 categories + neutral	0.536	/	/
Casel et al., 2021	En	feature-based and deep-learning based approaches	Subcorpus of the REMAN (literature) and TEC (tweets) corpora (3041 instances)	10 categories (Plutchik + Neutral + Other)	REMAN: 0.46 TEC: 0.53	4 categories	REMAN: 0.59 TEC: 0.57
Cortal et al., 2023	Fr	DistillCamemBERT	Guided narratives (3082 questionnaires answers)	4 categories	0.847	4 categories	0.931
Etienne et al., 2024	Fr	CamemBERT	Texts for children (5374 annotated sentences)	12 categories	0.42	4 categories	0.6
Noblet, 2025	Fr	GTE	Opinion posts on innovations (4980 sentences)	/	/	4 categories	0.635 (calculated by us)
Ours	Fr	CamemBERTaV2	Transcriptions of patients' discourse (7471 sentences)	8 categories	0.44 with 8 categories, 0.56 for the 6 main ones	4 categories	0.533 with 4 categories, 0.7 with the 3 main ones

Table 1: Comparison with similar works

cognitive parameters, such as the *experiencer*, *object*, and *target* of emotion (Fraisse and Paroubek, 2018), or appraisal-theoretic facets (Klinger, 2023; Troiano et al., 2023).

2.2 Automatic systems for emotion categories and expression mode detection

Table 1 above reviews recent corpora annotated for emotional categories, with particular attention to studies that also consider expression modes or elements of the component process model of emotions (a more detailed account of the annotation guidelines employed in these studies is provided in Appendix A.).

As shown, the corpora used for training in these tasks originate from diverse sources, including movie subtitles (OPUS corpus, Lison and Tiedemann, 2016), tweets and literature (TEC tweet corpus, Mohammad, 2012; REMAN literature corpus, Kim and Klinger, 2018), guided questionnaires obtained in the context of emotion regulation interventions, children's texts, and opinion posts on innovations. This diversity reflects the broad applicability of emotion analysis across domains.

Given that our study focuses on French-language corpora and that few systems are available, we primarily review previous work on the detection of emotional categories and expressive modes in French. Most recent approaches aiming for sentence-level classification rely on the BERT family of models and, in particular, its French variant CamemBERT (Martin et al., 2020). An exception is Noblet (2025), who employs the more recent GTE model (Zhang et al., 2024).

With respect to the granularity of emotion category annotation, most authors either (i) allow multiple emotions to be assigned to a single instance (multilabel classification) or, (ii) are constrained by their dataset design to assign only one label per example (multiclass classification). For instance, Öhman et al. (2020) and Etienne et al. (2024) adopt a multilabel approach, while Casel et al. (2021) apply multilabel classification on the REMAN dataset but switch to multiclass classification for the TEC corpus. Although multilabel approaches more accurately reflect the complexity and overlap of emotional expression (Plaza-del-Arco et al., 2024), they are computationally more challenging to implement. By contrast, multiclass setups are simpler to optimize, but they tend to oversimplify emotional nuances by enforcing mutual exclusivity among categories.

Regarding detection performance, reported results vary, with macro-F1 scores typically ranging from 0.42 to 0.84 for emotion categories and from 0.53 to 0.93 for expression modes or components in French and English. Performance also depends on the type of corpus and the number of annotated categories: the higher the number of categories in the annotation guidelines, the lower the inter-annotator agreement and evaluation scores tend to be (Bayerl and Paul, 2011). Cortal et al. (2023) achieved the highest

performance (macro-F1 0.84 for emotion categories and 0.93 for component detection) using a French corpus specifically designed around emotion components and limited to four emotion classes. By contrast, Etienne et al. (2024) reported more modest scores (0.42 and 0.6 respectively) on a pre-existing French corpus of children's texts annotated with a more fine-grained set of twelve emotion categories.

These studies pursue different application-oriented objectives such as detecting abusive language online (Öhman et al., 2020), analyzing text complexity in children's literature (Etienne et al., 2024), evaluating opinions on innovation-related ideas (Noblet, 2025), or more generally, examining the relationship between psychologically grounded emotion models and their linguistic realization (Cortal et al., 2023).

While these studies provide valuable insights, they remain limited in several respects. French corpora are still scarce and often highly domain-specific, multilabel emotion detection is also underexplored, and expression modes are defined inconsistently across studies. To address these gaps, our study introduces a systematically annotated corpus of French patient discourse, enriched with detailed information on modes of emotional expression, and provides a comparative evaluation of state-of-the-art LMs on the task of emotion classification. To our knowledge, this is the first study to perform sentence-level detection of both emotional categories and expression modes in direct patient discourse – the central focus of our research.

3. The GREMO-ABI project

Our study brings together linguistics, clinical psychology, and natural language processing (NLP) and is conducted within the participatory research project *Regulating Emotions and Behaviors After Brain Injury* (NCT 05 39 34 92). The project evaluates the effects of Dialectical Behavioral Therapy (DBT; Linehan, 2017) on individuals diagnosed with ABI. DBT is an intensive 15-month therapeutic program designed to improve emotion regulation through a combination of group sessions and individual psychotherapy. Patients with ABI often experience emotional dysregulation and alexithymia (Krasny-Pacini, 2020), conditions for which evidence-based interventions remain rare.

The overall aim of the GREMO-ABI project is to measure the effectiveness of DBT in patients with ABI. From a linguistic perspective, our goal is to develop and validate a novel approach for objectively assessing emotional regulation in patient discourse. This involves identifying formal linguistic and pragmatic markers that may serve as supplementary indicators of therapeutic progress.

Our approach extends this research trajectory by focusing on the implicit means of emotional expression. Drawing on psychological theories of emotion (Scherer et al., 2001) and clinical therapy frameworks (Linehan, 2017), we examine how emotion regulation might be captured through its linguistic realization. We hypothesize that an individual's capacity for emotional regulation is closely linked to the manner in which emotions are expressed in discourse. Detecting emotions and analyzing expression strategies may therefore serve as valuable indicators of emotion regulation. This approach could also provide a tool for identifying difficulties in emotional expression, such as alexithymia (Sifneos, 1996), in populations affected by ABI (Kuppelin et al., 2024; 2026), borderline personality disorder (Weiner, 2019), or autism spectrum disorder (Bemmouna et al., 2022).

4. Method

4.1 Corpus acquisition

In addition to the standard clinical outcome measures, semi-directive interviews were conducted with 37 patients with chronic ABI. Each interview lasted between 40 and 90 minutes.

Number of patients	Recor-dings	Duration	Patient's sentences	Patient's tokens
8	20	19h15min13s	7471	132,807

Table 2: Corpus data

Interviews were carried out at three time points, each spaced five months apart. During each session, the interviewer asked patients to recount emotionally salient memories and to describe their feelings in response to emotionally evocative or neutral images. The first interview took place at the beginning of the baseline phase (T0, five months before therapy). The second one was conducted immediately before therapy began (T1) to assess retest effects and response stability. The third session (T2) occurred immediately after the end of therapy and aimed to evaluate potential gains in emotion regulation following five months of intensive DBT. All participants were fluent speakers but exhibited impaired cognitive functions and pathological scores on clinical evaluation scales measuring emotional regulation difficulties (Klein et al., 2024; Klein, 2025).

We analyze a subset of the data, comprising 20 voice-recorded interviews with eight patients across the three time points. Some T0 recordings were unavailable due to scheduling constraints. Table 2 provides an overview of the dataset. The recordings were automatically transcribed using the Whisper tool (Bain et al., 2023; Radford et al., 2023), which provides sentence segmentation and identification of the different participants in the exchange. It also aligns the text with the audio

```
<pat id="006_A-50"> Donc euh, le lycée quand j'étais euh à <npr/>. <seg time= "358"/></pat>
<pat id="006_A-51"> Parce que c'est vrai qu'il y avait beaucoup <rep> beaucoup beaucoup
</rep> de <rep> de euh de </rep> d'in- pas d'injustice, mais de euh de d'inégalités, en
fait, entre guillemets euh <seg time= "370"/></pat>
<psy id="006_A-52a"> vis-à-vis de ? </psy>
```

Figure 1: Corpus example with unique ids, disfluences and time codes

signal, preserving temporal information. At this stage, a unique identifier is assigned to each sentence, enabling the corpus to be reshuffled for out-of-context annotation[3]. The entire corpus is then manually reviewed to ensure anonymization and to correct transcription errors.

As part of the project's interdisciplinary design, transcripts were further enriched with information on hesitations and disfluencies to facilitate cross-analysis with phonological properties (see Briand et al., 2022 and Forth.). Figure 1 illustrates a sample of the corpus in its final pre-annotation state, showing unique sentence IDs, turn-taking labels ("<psy>" and "<pat>"), anonymisation ("<npr/>"), disfluency annotations ("<rep>" for repetitions) and temporal segmentation ("<seg time=…/>").

4.2 Annotation guidelines and process

Our annotation scheme (described in detail in Klein et al., 2024; 2026) is inspired by Plantin (2011; 2014), Micheli (2014), Etienne and Battistelli (2021), and the DBT model of emotional processes (Linehan, 2017; Weiner et al., 2022). It is illustrated in Appendix B. We opt for a categorical approach, which classifies emotions into discrete labels, as it offers clearer interpretability and aligns more closely with everyday emotional concepts, as well as with the labels explicitly taught to patients during therapy. While dimensional models (§ 2.1) capture subtle shifts in affect, categorical models facilitate the identification of specific emotional patterns.

Within this approach, the annotation procedure is divided into two subtasks:

- Sentence-level annotation: each sentence is labeled for *emotionality* (yes/no), *polarity* (positive, negative, mixed, or neutral/uncertain), *intensity* (intense vs. non-intense), and *emotional category* (anger, disgust, embarrassment, fear, jealousy, joy, love, sadness). The typology of categories was derived from the emotion set central to DBT. Each emotional category acts as an umbrella term for semantically related terms. For example, *anger* also encompasses *frustration*, *irritation*, and *rage*. Emotional category annotation is multilabel, with annotators allowed to assign up to two emotions per sentence, thus providing a more realistic account of emotional expression (Demsky et al., 2020; Koufakou and Nieves, 2025).
- Token-level annotation: in sentences marked as emotional, the lexical items that triggered the emotional interpretation were annotated for expression mode (*Designated*, *Suggested*, *Manifested*, or *Uncertain*).

The final annotation guidelines and category inventory were consolidated through consultation with annotators and iterative revisions. The annotation team comprised 13 participants, all of them in their early twenties. They were undergraduate or graduate students in linguistics or speech and language therapy, recruited as part of an internship. They attended a half-day training session and were able to consult project members throughout the annotation campaign whenever questions arose.

Annotations were carried out on a sentence-by-sentence basis, without access to the broader discourse context, using the INCEpTION platform (Klie et al., 2018). An example of sentence- and token-level annotation in INCEpTION is provided in Appendix C. All annotations were subsequently reviewed and harmonized to ensure consistency and quality control.

Inter-rater agreement was calculated using Cohen's kappa (1960). To this end, half of the corpus was independently re-annotated by a second team of nine annotators. The average agreement scores were 0.58 for emotionality, 0.55 for polarity (four classes), 0.45 for intensity (two classes), 0.37 for emotion category (eight classes), and 0.32 for emotion modes.

Notably, intensity annotation proved less consensual than polarity, despite being coded as a binary distinction (Intense vs. Non-Intense). This may reflect the multiple ways intensity can be conveyed in speech, either intrinsically through lexical items (e.g., *depressed*, *incredible*) or extrinsically through intensifiers (e.g., *great*, *totally*). The relatively higher agreement for emotion categories may be related to the broader semantic distinctions provided by the larger set of

[3]The annotation is performed in random order to minimize potential empathization effects; however, the released annotated corpus will preserve the original conversational sequence.

classes. In contrast, the low agreement score for modes of expression suggests that these categories remain difficult to operationalize reliably, even for human annotators.

Agreement in unit segmentation was assessed using Krippendorff's alpha (Hayes and Krippendorff, 2007). The resulting score of 0.41 suggests that annotators may have found it easier to identify segments conveying emotional information than to consistently assign them to one of the three proposed categories: Designated, Suggested, or Manifested.

5. Annotation results

An overview of the annotations is presented in Table 3.

Task	Properties	Number	Proportion
Emotionality	Yes	3,812	51.02%
	No	3,659	48.98%
Polarity	Positive	1,230	32.27%
	Negative	1,906	50.00%
	Mixed	369	9.68%
	Neutral/Uncertain	307	8.05%
Intensity	Intense	813	21.33%
	Non-intense	2,999	78.67%
Emotional category	Anger	612	13.95%
	Disgust	392	8.94%
	Embarrassment	291	6.63%
	Fear	547	12.47%
	Jealousy	48	1.09%
	Joy	1,204	27.45%
	Love	438	9.99%
	Sadness	854	19.47%

Table 3: Annotation overview at sentence-level

5.1 Sentence-level annotations

The dataset is relatively balanced in terms of emotionality: 51.02% of sentences were annotated as emotional, while the remaining 48.98% were labeled non-emotional. All subsequent annotation layers were therefore applied to the 3,812 emotional sentences.

With respect to polarity, half of the emotional sentences were negative (50.00%), while nearly one third were positive (32.27%). Mixed polarity occurred in 9.68% of cases, and 8.05% were annotated as *neutral* or *uncertain*. This distribution confirms that negative emotions are slightly more prevalent, while a significant proportion of utterances convey complex or ambiguous evaluative content.

In terms of intensity, the majority of emotional sentences were annotated as non-intense (78.67%), with only 21.33% marked as intense. This suggests that strongly expressed emotions represent a minority in patient discourse, though they constitute a clinically relevant subset for therapeutic monitoring.

Finally, regarding emotion categories, the most frequently expressed emotions were *joy* and *sadness*, likely reflecting their status as prototypical representatives of positive and negative valence, respectively. These were followed by *anger* (13.95%), *fear* (12.47%) and *love* (9.99%). The relatively lower frequencies of certain categories such as *jealousy*, *embarrassment* and *disgust* can be explained by the fact that such emotions are often perceived as less "noble" (i.e. socially acceptable) and are therefore less likely to be explicitly verbalized in interview contexts.

The total number of annotations for the emotional category feature (3,505) exceeds the number of emotional sentences (3,812 minus the neutral/uncertain cases annotated without a precise emotional category) because the annotation scheme allowed for double labeling, thereby capturing instances in which multiple emotions were expressed within a single sentence.

5.2 Token-level annotations

A total of 6,260 spans of emotional expressions were annotated across the 3,812 emotional sentences in the corpus, representing an average of just over 1.6 emotional expressions per sentence. Of these expressions, 28.12% were annotated as Designated, meaning they directly express an emotional feeling, as in *angry*, *I couldn't take it anymore* or *I was getting angry*. 51.47% of emotional expressions were labeled as Suggested, referring to emotion-evoking elements likely to trigger affective responses, such as *problem*, *it's very hard*, and *I don't have time to do anything anymore*. Manifested expressions accounted for 19.57%, denoting behavioral or physiological consequences of an emotion, such as *wanting to kill myself*, *smiling*, and *ruminating*. The remaining 0.84% were annotated as Uncertain, reflecting the frequent occurrence of ill-formed or ambiguous multiword expressions typical of spontaneous speech. The average length of the emotional expression is 1.96 tokens, with 3,030 consisting of a single token, such as *stress*, *difficult*, and *tired*.

Overall, the annotation scheme yielded a rich and diversified distribution of emotional features, confirming the feasibility of large-scale annotation of patient discourse while also highlighting the inherent variability and ambiguity of emotional expression in spoken interaction.

6. Classification experiments

6.1 Preprocessing

The corpus underwent minimal preprocessing. While XML tags were removed, hesitations and disfluencies (e.g., repetitions, revisions) were retained to preserve syntactic and prosodic cues. Each sentence's annotation was then encoded as a binary vector representation, which was subsequently sliced to generate the target labels required for training. Phonological enrichments were also retained for the sentence-level classification experiments, as they may provide pragmatic and prosodic cues relevant to emotionality.

The classification of emotional expression modes was performed at the sentence level by converting the annotations into a multi-label classification task, following the approaches of Cortal et al. (2023) and Etienne et al. (2024). A more fine-grained, intra-sentential classification will be addressed in future work.

For model training, the processed sentences served as input to transformer-based architectures, primarily CamemBERT and its variants, alongside multilingual baselines. All models were implemented in Python using the Hugging Face Transformers library (Wolf et al., 2020). The corpus was split into training, validation and test sets (70/10/20% respectively). Each sentence was tokenized according to the subword vocabulary of the corresponding model, and the resulting embeddings were used for binary (emotionality, intensity) and multilabel (polarity, emotional category, expression mode) sentence classification tasks. In what follows, we report results for the annotation of emotionality, emotional category, and expression mode classification. We trained all the models for 10 epochs with a learning rate of 5^{e}-5 and a batch size of 16.

6.2 Considering all the categories

We first conducted a single training run with each of the seven models: CamemBERT, CamemBERTav2, DistillCamemBERT, mDeBERTav3, and GTE, a pre-trained DistillCamemBERT for emotion and component detection (Cortal et al., 2023) and a pre-trained CamemBERT (see footnote 1) also trained for emotion detection (Ekman's typology + neutral). All models were fine-tuned across all layers, except for the GTE model, whose size made full fine-tuning impractical. Following Noblet (2025), all layers of the GTE model were frozen except the final one, which was fine-tuned on our data. Each model was evaluated both in its fine-tuned configuration and as a baseline, using logistic regression for comparison (see shaded lines in Tables 4 and 5).

For the binary classification task of emotionality, the Cortal et al. (2023) model achieved the highest performance, reaching an F1 score of 0.876 after additional fine-tuning on our corpus. This configuration outperformed the original DistillCamemBERT backbone, suggesting a degree of transfer learning between related annotation frameworks. Nonetheless, all models performed fairly similarly on this simpler binary task, with F1 scores ranging from 0.85 to 0.87 after fine-tuning.

Model	Emotio-nality	Emotion category		Expression mode	
		Micro F1	Macro F1	Micro F1	Macro F1
C	0.865	0.338	0.133	0.716	0.527
C2	0.868	0.569	0.408	**0.720**	**0.533**
CD	**0.876**	0.567	0.43	0.687	0.511
AC	0.868	0.508	0.316	0.687	0.503
D	0.866	**0.573**	0.43	0.695	0.514
M	0.858	0.537	0.387	0.681	0.505
G	0.851	0.518	**0.44**	0.586	0.416
C + LR	0.845	0.335	0.233	0.521	0.361
C2 + LR	0.769	0.249	0.150	0.437	0.305
CD + LR	0.855	0.490	0.385	0.612	0.450
AC + LR	0.843	0.479	0.319	0.567	0.400
D + LR	0.86	0.434	0.313	0.565	0.401
M + LR	0.719	0.031	0.016	0.196	0.095
G + LR	0.825	0.499	0.406	0.590	0.435

Table 4: F1 micro and macro scores for emotionality, emotion category and expression mode detection.
C = CamemBERT, C2 = CamemBERTav2,
D = DistilCamemBERT, CD = Cortal DistillCamemBERT,
M = mDeBERTa, G = GTE, AC = Astrosbd's CamemBERT,
LR = Logistic Regression

In terms of modes of emotional expression, CamemBERTav2 achieved the best results, despite showing a lower baseline score than other models. Considering the emotion category classification, performance patterns were less consistent: the best-performing model varied depending on whether micro- or macro-averaged F1 scores were taken into account. These findings indicate differential strengths across

models, with some excelling in the detection of specific emotion categories, while others demonstrate more robust, generalized performance. Interestingly, Astrosbd's CamemBERT systematically underperforms compared to Cortal's, suggesting that its training data differ even more from ours.

This initial evaluation provided a baseline estimate of model performance, allowing us to exclude low-frequency categories and discard underperforming models in subsequent experiments (see Table 4).

6.3 Removing minor categories

We then re-trained the remaining models three times, using different random seeds to assess performance variability across runs and identify the best-performing configuration for each task (Table 5). To minimise the impact of underrepresented categories on overall performance, we removed emotion and expression categories with very few annotations. For emotion categories, the labels *Embarrassment* (6.63%) and *Jealousy* (1.09%) were removed. Regarding modes of emotional expression, the *Uncertain* category was also removed, as it accounted for only 0.84% of annotations. Given that these suppressed emotional categories are infrequently verbalized and are associated with a relatively limited vocabulary, we chose to exclude them rather than artificially inflate their frequency in an attempt to balance the dataset.

Model	Emotion category		Expression mode	
	Micro F1	Macro F1	Micro F1	Macro F1
C	0.53	0.42	**0.72**	**0.70**
C2	**0.62**	**0.58**	0.70	0.69
CD	0.60	0.56	0.71	0.69
G	0.53	0.45	0.58	0.55
C + LR	0.35	0.30	0.52	0.48
C2 + LR	0.26	0.19	0.44	0.41
CD + LR	0.5	0.46	0.62	0.60
D + LR	0.45	0.39	0.57	0.54
G + LR	0.52	0.48	0.60	0.58

Table 5: F1 scores, averages over 3 runs, without lesser-annotated categories

In terms of emotional category classification, and with the new version of the datasets, the CamemBERTaV2 model achieved the best performance, with a macro F1 score of 0.58. However, the double-fine-tuned model by Cortal et al. (2023) performed only marginally worse. Interestingly though, the best-performing models differed from those identified in the previous experiment once minority categories were excluded, indicating that class imbalance exerts a measurable influence on model ranking.

A confusion matrix was used to analyze the most ambiguous emotion categories for the model. Figure 2 displays the CamemBERTaV2 model's results on the test corpus. The model most accurately recognized *joy*, with a recall rate of 75 %, while *disgust* achieved the lowest recognition rate (34%). This outcome is not surprising, given that *disgust* remains one of the least represented categories (8.94%; see Table 3) even after the removal of minority labels.

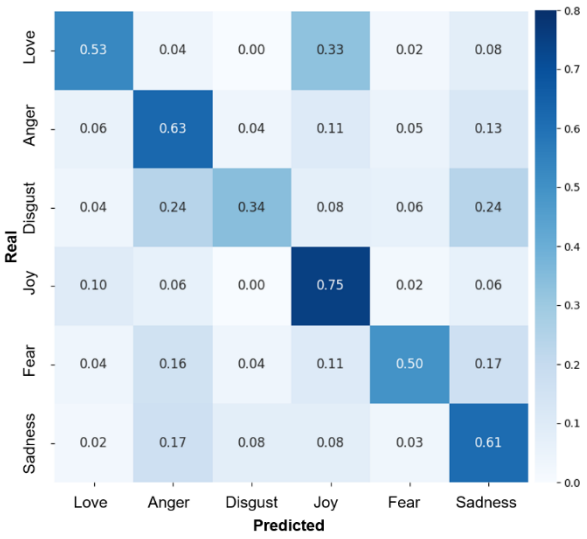


Figure 2: Real and predicted emotional categories on test split with CamemBERTav2

Although *disgust* originates from Ekman's typology, it is known to be rarely verbalized and difficult to detect linguistically (see Alm et al., 2005 and Alm, 2010 who merge it with *anger*). Consequently, our model frequently confuses *disgust* with *sadness* and *anger* (24% each).

Regarding the detection of emotional expression modes, it should be noted that the CamemBERT model, previously outperformed by its v2 variant when all categories were included, achieved slightly higher scores on this specific detection task. The other two models, CamemBERTav2 and DistillCamemBERT (Cortal et al., 2023), reached comparable levels of performance.

As for the baseline results, the Cortal model (CD + LR) consistently outperformed nearly all other baselines across tasks, particularly when compared to the DistillCamemBERT backbone model baseline (D + LR). This finding further supports the hypothesis of transferability between models trained on related emotional properties.

6.4 Comparison with generative models

In order to compare the models trained on manually annotated data, we conducted an experiment by generating annotations on the test split (1493 sentences) with two generative models locally deployed with Ollama[4] (Lin and Safi, 2025): Llama3.2 (3B parameters) and Llama3.1 (8B parameters) (Touvron et al., 2023). We adopted a zero-shot approach by providing the models with definitions of the annotation categories. However, we deliberately refrained from supplying annotated examples from the corpus in order to preserve a true baseline for assessing the generative model's intrinsic capacities. The prompts (in French) are provided in Appendix D. Evaluation metrics are reported in Table 6 below.

[4] https://ollama.com/

Models		Emotion-ality	Emotion category	Expression mode
Llama3.2 (3B)	P	0.64	0.128	0.17
	R	0.48	0.189	0.259
	F1	0.316	0.153	0.206
Llama3.1 (8B)	P	0.687	0.127	0.279
	R	0.674	0.152	0.39
	F1	0.671	0.138	0.325
Ours (fine-tuned Camembertav2)	P	0.87	0.656	0.669
	R	0.872	0.597	0.801
	F1	**0.872**	**0.625**	**0.729**

Table 6: Evaluation metrics for comparison of annotation on test split

Interestingly, the larger Llama3.1 model outperforms the smaller version on the *emotionality* and *expression mode* features, but not on the more complex *emotion category* feature. Nevertheless, our fine-tuned model achieves superior performance across all tasks compared with the generative models.
These scores confirm that our fine-tuned model outperforms a minimally trained generative model. The most substantial gain is observed in the *emotion category* feature, where the generative model achieves a maximum score of 0.153, compared with 0.625 for our model.

7. Discussion

Compared with a closely related task of expression mode detection, Cortal et al. (2023) report strong results for component detection, achieving a macro F1 score of 0.93 with CamemBERT, whereas our best score reached 0.70. This discrepancy can likely be attributed to differences in corpus design: Cortal's et al. dataset consists of self-annotated questionnaires containing sections that explicitly target certain emotion components. Consequently, their data are probably denser in emotion-related content than our patient discourse corpus. Moreover, their dataset is built from written texts, while our dataset is transcribed spoken language. Nevertheless, we observed that transfer learning from their pretrained model proved partially effective, improving both emotion category and expression mode recognition, and providing strong baseline performance.

In parallel, Etienne et al. (2024) report macro F1 scores of 0.42 for emotion category detection and 0.64 for expression mode task. Despite several differences in annotation schemes, these results are consistent with our own (0.58 and 0.70, respectively). Reducing the number of categories in our experiments might have led to higher scores. However, we believe that any such gains would likely have been offset by the model's reduced capacity to handle non-standard and disfluent sentence structures, which are characteristic of our data.

Overall, our findings corroborate those of Etienne et al. (2024), confirming that this range of evaluation scores is typical for current NLP techniques for recognizing emotions. Despite the additional complexity of working with spoken discourse transcriptions, our annotation framework and classification models achieve good performance within this challenging domain, especially when compared to the performances of a zero-short generative model.

Comparisons with other studies still remain very difficult. The emotional categories and typologies of modes of expression are specific to each study, which makes the detection scores comparable, but prevents more precise evaluation without compromising the integrity of our annotation scheme. Moreover, the categories may vary in definition depending on the typology adopted. Also, most previous studies have been conducted on written texts, whereas our corpus exhibits significant features of spoken language. To the best of our knowledge, no comparable French spoken datasets annotated for emotions in such detailed manner are currently available for comparison with similar datasets.

8. Conclusion and future work

This paper introduced an annotation framework that integrates insights from previous psychological and linguistic research on emotion. Applying this annotation scheme enabled the development of a competitive model for the detection of emotion categories and modes of expression, fine-tuned on a corpus of transcribed French patient discourse.

Despite the added challenges of working with non-standard written text derived from spoken interaction, our model outperformed non-fine-tuned baselines and achieved competitive results compared with similar emotion recognition tasks. These findings highlight the potential of NLP tools to support the linguistic assessment of emotional regulation in clinical contexts. Thus, beyond methodological implications, this work could serve as a complementary tool for clinicians and it opens promising avenues for applied research.

Future work will extend this approach to span-level analysis and prosodic features, providing deeper insight into how emotion is expressed, regulated, and transformed through language in therapeutic interaction.

Limitations

Several contextual factors may have influenced the recordings and annotations. These include the interviewer-interviewee relationship (e.g., familiarity, prior sessions), recording order, individual patient circumstances (such as relapses or life events), and variations in interviewer prompting styles. Such variables are inherent to naturalistic clinical data collection but should be considered when interpreting the results.

Aknowledgements

This work of the Interdisciplinary Thematic Institute LIRIC, as part of the ITI 2021-2028 program of the University of Strasbourg, CNRS and Inserm, was supported by IdEx Unistra (ANR-10-IDEX-0002), and by SFRI-STRAT'US project (ANR-20-SFRI-0012) under the framework of the French Investments for the Future Program.

This study was also supported by LiLPa (UR 1339) from the University of Strasbourg (France) and uses data from the GREMO-ABI project supported by UGECAM Alsace and by the French Eastern Interregional Group of Clinical Research and Innovation (GIRCI Est; APJ 2021). We wish to thank EMOI-TC patients for their participation in the protocol and for sharing their emotional and behavioral difficulties in the recording.

We are also grateful to Lydie NGWEN-MBANG, Master student in Language Technology at the University of Strasbourg. Her work during the internship in March 2025 served as a basis to design the prompts for the comparison of our system with llama models.

Ethics Statement

The study is referenced as NCT: 05 39 34 92 in Clinical Trials and was approved by the research board CPP ("protection committee") of Ile de France XI (number: ID-RCB 2021-A01996-35). Procedures complied with the ethics code outlined in the Declaration of Helsinki.

References

Alm, C. (2010). Characteristics of high agreement affect annotation in text. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 118–122, Uppsala, Sweden. Association for Computational Linguistics.

Alm, C. O., Roth, D., and Sproat, R. (2005). Emotions from Text: Machine Learning for Text-based Emotion Prediction. In Mooney, R., Brew, C., Chien, L.-F., and Kirchhoff, K. (Eds.), *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing,* pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Aman, S. an Szpakowicz, S. (2007). Identifying Expressions of Emotion in Text. *In Proceedings of the 10th International Conference Text, Speech and Dialogue*, pages 196–205, Pilsen, Czech Republic.

Antoun, W., Kulumba, F., Touchent, R., de la Clergerie, É., Sagot, B., and Seddah, D. (2024). CamemBERT 2.0: A Smarter French Language Model Aged to Perfection. arXiv:2411.08868 [cs].

Augustyn, M. (2015). Annotations des marques de la subjectivité langagière : discours rapporté, passages entre guillemets et lexique des affects. *Manuel de codage.*

Augustyn, M., Hamou, S., Bloquet, G., Goossens, V., Loiseau, M., and Rinck, F. (2008). Lexique des affects : constitution de ressources pédagogiques numériques. In Loiseau M. et al. (Eds.), *Proceedings of the Colloque International des Étudiants Chercheurs en Didactique des Langues et en Linguistique*, pages 407–414, Grenoble, France. Presses Universitaires de Grenoble.

Azam, N., Ahmad, T., and Ul Haq, N. (2021). Automatic emotion recognition in healthcare data using supervised machine learning. *PeerJ Computer Science*. 2021 Dec 7, e751.

Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *Interspeech*.

Bayerl, P. S. and Paul, K. I. (2011). What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation. *Computational Linguistics*, 37(4):699–725.

Bemmouna, D., Coutelle, R., Weibel, S., & Weiner, L. (2022). Feasibility, Acceptability and Preliminary Efficacy of Dialectical Behavior Therapy for Autistic Adults without Intellectual Disability: A Mixed Methods Study. *Journal of Autism and Developmental Disorders*, 52(10), 4337–4354.

Bertolini, L., Elce, V., Michalak, A., Widhoelzl, H.-S., Bernardi, G., and Weeds, J. (2024). Automatic Annotation of Dream Report's Emotional Content with Large Language Models. In Yates, A., Desmet, B., Prud'hommeaux, E., Zirikly, A., Bedrick, S., MacAvaney, S., Bar, K., Ireland, M., and Ophir, Y. (Eds.), *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 92–107, St. Julians, Malta. Association for Computational Linguistics.

Briand, T., Fauth, C., and Vassiliadou, H. (2022). Marques de l'émotion dans la fluence d'un patient cérébrolésé : Étude préliminaire de faisabilité. In *XXXIVe Journées d'Études sur la Parole -- JEP 2022*, pages 90–98, Île de Noirmoutier, France. ISCA.

Briand, T., Vassiliadou, H., Fauth, C., Klein, S., Todirascu, A., Kuppelin, M., and Krasny-Pacini, A. (Forth.). L'annotation de l'expression des émotions dans les récits des patients atteints de lésions cérébrales acquises : méthodes, difficultés et défis pour le traitement des corpus oraux. *Corela* Special Issue.

Buechel, S. and Hahn, U. (2017). EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In Lapata, M., Blunsom, P., and Koller, A. (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Casel, F., Heindl, A., and Klinger, R. (2021). Emotion Recognition under Consideration of the Emotion Component Process Model. In Evang, K., Kallmeyer, L., Osswald, R., Waszczuk, J., and Zesch, T. (Eds.), *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 49–61, Düsseldorf, Germany. KONVENS 2021 Organizers.

Chen, S.-Y., Hsu, C.-C., Kuo, C.-C., Huang, T.-H., and Ku, L.-W. (2018). EmotionLines: An Emotion Corpus of Multi-Party Conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chen, Y., Lee, S. Y. M., and Huang, C.-R. (2009). A cognitive-based annotation system for emotion computing. In *Proceedings of the Third Linguistic Annotation Workshop on - ACL-IJCNLP '09*, pages 1–9, Suntec, Singapore. Association for Computational Linguistics.

Chiril, P., Pamungkas, E. W., Benamara, F., Moriceau, V., and Patti, V. (2022). Emotionally Informed Hate Speech Detection: A Multi-target Perspective. *Cognitive Computation*, 14(1):322–352.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 37–46.

Cortal, G., Finkel, A., Paroubek, P., and Ye, L. (2023). Emotion Recognition based on Psychological Components in Guided Narratives for Emotion Regulation. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 72–81, Dubrovnik, Croatia. Association for Computational Linguistics.

Cortal, G. (2024). Sequence-to-Sequence Language Models for Character and Emotion Detection in Dream Narratives. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14717–14728, Torino, Italia. ELRA and ICCL.

Delestre, C. and Amar, A. (2022). DistilCamemBERT : une distillation du modèle français CamemBERT. In *CAp (Conférence sur l'Apprentissage automatique)*, Vannes, France.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Dragos, V., Battistelli, D., Etienne, A., and Constable, Y. (2022). Angry or Sad ? Emotion Annotation for Extremist Content Characterisation. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S. (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 193–201, Marseille, France. European Language Resources Association.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3–4):169–200.

Étienne, A. and Battistelli, D. (2021). Annotation manuelle des émotions dans des textes écrits avec la plateforme Glozz. Technical report, MoDyCo ; Université Paris Nanterre.

Etienne, A., Battistelli, D., and Lecorvé, G. (2024). Emotion Identification for French in Written Texts: Considering Modes of Emotion Expression as a Step Towards Text Complexity Analysis. In De Clercq, O., Barriere, V., Barnes, J., Klinger, R., Sedoc, J., and Tafreshi, S. (Eds.), *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 168–185, Bangkok,

Thailand. Association for Computational Linguistics.

Fraisse, A. and Paroubek, P. (2014). Toward a unifying model for Opinion, Sentiment and Emotion information extraction. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S. (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC`14)*, pages 3881–3886, Reykjavik, Iceland. European Language Resources Association (ELRA).

Fraisse, A. and Paroubek, P. (2018). Opinion/Sentiment/Emotions annotations. Online annotation guide. Available at: https://perso.limsi.fr/pap/DEFT2018/annotation_guidelines/index.html

Giouli, V., Fotopoulou, A., and Mouka, E. (2014). Annotating sentiment expressions for lexical resources. In Blumenthal, P., Novakova, I., and Siepmann, D. (Eds.), *Les émotions dans le discours / Emotions in Discourse*, Peter Lang D, pp. 281–296.

Goossens, V. (2015). Les noms d'affect parmi les noms abstraits intensifs : nouvelles perspectives typologiques. *Langue française*, 185(1):59–72.

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1), 77–89.

He, P., Liu, X., Gao, J., and Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *Proceedings of the International Conference on Learning Representations (ICLR) 2021*.

Kim, E. and Klinger, R. (2018). Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions. In Bender, E. M., Derczynski, L., and Isabelle, P. (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kim, E. and Klinger, R. (2019). An Analysis of Emotion Communication Channels in Fan-Fiction: Towards Emotional Storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 56–64, Florence, Italy. Association for Computational Linguistics.

Klein, S. (2025). Proposition d'un cadre méthodologique pour l'étude de l'expression linguistique des émotions. In Kananovich, A., Belem, R. Y., Lacassain, C., Marsac, F., Vaxelaire, B., and Kleiber, G. (Eds.), *Regards sur la perception : de l'expérience au linguistique*, Recherches en PArole, Editions du CIPA, Mons, France, pp. 53–72.

Klein, S., Todirascu, A., Vassiliadou, H., Kuppelin, M., Becart, J., Briand, T., Coridon, C., Gerhard-Krait, F., Laroche, J., Ulrich, J., and Krasny-Pacini, A. (2024). Annotating Emotions in Acquired Brain Injury Patients' Narratives. In Demner-Fushman, D., Ananiadou, S., Thompson, P., and Ondov, B. (Eds.), In Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024, pages 26–36, Torino, Italia. ELRA and ICCL.

Klein, S., Benninger, C., Briand, T., Fauth, C., Gerhard-Krait, F., Kuppelin, M., Lammert, M., Laroche, J., Raguenet, L., Schnedecker, C., Todirascu, A., Krasny-Pacini, A., Vassiliadou, H. (2026). *Projet GREMO-LING, Annotation manuelle de l'expression de l'émotion dans des transcriptions de l'oral. Guide d'annotation pour les transcriptions de discours des patients présentant une Lésion Cérébrale Acquise (LCA).* https://hal.science/hal-05536247

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., and Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In Zhao, D. (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Klinger, R. (2023). Where are We in Event-centric Emotion Analysis? Bridging Emotion Role Labeling and Appraisal-based Approaches. In *Proceedings of the Big Picture Workshop*, pages 1–17, Singapore. Association for Computational Linguistics.

Koufakou, A. and Nieves, E. (2025). Review of recent emotion-annotated text corpora and resources. *Language Resources and Evaluation*.

Krasny-Pacini, A. (2020). Feasibility of an emotion regulation intervention based on dialectical behavior therapy (DBT) in adults with a chronic acquired brain injury. *WFNR/SOFMER 2020*, Oct 2020, Lyon, France. WFNR, 2020.

Kuppelin, M., Goetsch, A., Choisel, R., Isner-Horobeti, M.-E., Goetsch, T., & Krasny-Pacini, A. (2024). An exploratory study of dialectical behaviour therapy for emotional dysregulation and challenging behaviours after acquired brain injury. *NeuroRehabilitation*, 55(1), 77–94.

Kuppelin, M., Bemmouna, D., Weiner, L., Goetsch, T., & Krasny-Pacini, A. (2026). Emotion Dysregulation in Adults with Acquired Brain Injury: Conceptualization of Emotion Dysregulation, Validation of the French DERS-

16 Scale and its Utility in Clinical Practice. *NeuroRehabilitation*.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Kondrak, G. and Watanabe, T. (Eds.), *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Lin, H., & Safi, T. (2025). ollamar: An R package for running large language models. *Journal of Open Source Software*, 10(105), 7211.

Linehan, M. (2017). *Manuel d'entraînement aux compétences TCD.* Médecine et hygiène, Chêne-Bourg, 2e édition.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Lee, L.-H., Li, J.-H. and Yu, L.-C. (2022). Chinese EmoBank: Building Valence-Arousal Resources for Dimensional Sentiment Analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(4): 1–18.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., De La Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Mathieu, Y. Y. (2000). *Les verbes de sentiment*. CNRS Editions.

Micheli, R. (2014). *Les émotions dans les discours : modèle d'analyse et perspectives empiriques.* De Boeck Duculot, Louvain-la-Neuve, 1re édition.

Mohammad, S. (2012). #Emotional Tweets. In Agirre, E., Bos, J., Diab, M., Manandhar, S., Marton, Y., and Yuret, D. (Eds.), *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Mohammad, S. M. and Kiritchenko, S. (2015). Using Hashtags to Capture Fine Emotion Categories from Tweets. *Computational Intelligence*, 31(2):301–326.

Muhammad, S. H., Ousidhoum, N., Abdulmumin, I., Yimam, S. M., Wahle, J. P., Ruas, T. L., Beloucif, M., De Kock, C., Belay, T. D., Ahmad, I. S., Surange, N., Teodorescu, D., Adelani, D. I., Aji, A. F., Ali, F. D. M., Araujo, V., Ayele, A. A., Ignat, O., Panchenko, A., Zhou, Y. and Mohammad S. (2025). SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2558–2569, Vienna, Austria. Association for Computational Linguistics.

Noblet, J. (2025). Annotation et modélisation des émotions dans un corpus textuel : une approche évaluative. In *French language. 27{è}me Rencontre des {É}tudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL).* Marseille, France.

Ohman, E., Bizzoni, Y., Feldkamp Moreira, P., and Nielbo, K. (2024). EmotionArcs: Emotion Arcs for 9,000 Literary Texts. In Bizzoni, Y., Degaetano-Ortlieb, S., Kazantseva, A., and Szpakowicz, S. (Eds.), *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 51–66, St. Julians, Malta. Association for Computational Linguistics.

Öhman, E., Pàmies, M., Kajava, K., and Tiedemann, J. (2020). XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In Scott, D., Bel, N., and Zong, C. (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Paroubek, P., Grouin, C., Bellot, P., Claveau, V., Eshkol-Taravella, I., Fraisse, A., Jackiewicz, A., Karoui, J., Monceaux, L., and Torres-Moreno, J.-M. (2018). DEFT2018 : recherche d'information et analyse de sentiments dans des tweets concernant les transports en Île de France. In *Actes de la conférence TALN. Volume 2 - démonstrations, articles des rencontres jeunes chercheurs, ateliers DeFT*, pages 219–230, Rennes, France. Association pour le Traitement Automatique des Langues.

Pearl, L. and Steyvers, M. (2010). Identifying Emotions, Intentions, and Attitudes in Text Using a Game with a Purpose. In Inkpen, D. and Strapparava, C. (Eds.), *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of*

Emotion in Text, pages 71–79, Los Angeles, CA. Association for Computational Linguistics.

Plantin, C. (2011). *Les bonnes raisons des émotions*. Peter Lang CH.

Plantin, C. (2014). Construire, justifier, signifier, gérer l'émotion en interaction : les opérations de construction des séquences émotionnées. In Pustka E., Goldschmitt S. (Eds.), *Emotionen, Expressivität, Emphase, Studienreihe Romania*. Erich Schmidt Verlag, Berlin.

Plaza-del-Arco, F. M., Curry, A., Cercas Curry, A., and Hovy, D. (2024). Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC/COLING 2024)*, pages 5696–5710, Torino, Italy. ELRA and ICCL.

Plutchik, R. (1991). *The emotions*. University Press of America, Lanham, Md, Rev. ed.

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., and Shulman, E. (2016). Modelling Valence and Arousal in Facebook posts. In Balahur, A., van der Goot, E., Vossen, P., and Montoyo, A. (Eds.), *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, pages 28492–28518, Virtual. PMLR.

Ray, A., Mishra, S., Nunna, A., and Bhattacharyya, P. (2022). A Multimodal Corpus for Emotion Recognition in Sarcasm. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S. (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6992–7003, Marseille, France. European Language Resources Association.

Roman, N. T., Piwek, P., Carvalho, A. M. B. R., and Alvares, A. R. (2015). Sentiment and Behaviour Annotation in a Corpus of Dialogue Summaries. *Journal of Universal Computer Science*, 21(4):561–586.

Russell, J. A. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

Scherer, K. R. (2001), Schorr, A., and Johnstone, T. (Eds.). Appraisal processes in emotion: theory, methods, research. *Series in Affective Science*. Oxford University Press, Oxford, New York.

Scherer, K. R. and Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310–328.

Sifneos, P. E. (1996). Alexithymia: Past and present. *The American Journal of Psychiatry*, 153(Suppl):137–142.

Sosea, T. and Caragea, C. (2020). CancerEmo: A Dataset for Fine-Grained Emotion Detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904, Online. Association for Computational Linguistics.

Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing - SAC '08*, page 1556-1560, Fortaleza, Ceara, Brazil. ACM Press.

Tessore, J. P., Esnaola, L. M., Lanzarini, L., and Baldassarri, S. (2022). Distant Supervised Construction and Evaluation of a Novel Dataset of Emotion-Tagged Social Media Comments in Spanish. *Cognitive Computation*, 14(1):407–424.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2302.13971

Troiano, E., Oberländer, L., and Klinger, R. (2023). Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction. *Computational Linguistics*, 49(1):1–72.

Vidrascu, L. (2007). *Analyse et détection des émotions verbales dans les interactions orales*. Ph.D. thesis, Université Paris Sud - Paris XI.

Weiner, L. (2019). Évaluation de la faisabilité et de l'efficacité d'un groupe de thérapie comportementale dialectique (TCD) transnosographique : le groupe de régulation EMOtionnelle (GREMO). *French Journal of Psychiatry*, 1:S24–S25.

Weiner, L., Bemmouna, D., Weibel, S., Lachaux, E., Derrouazi, S., Poussardin, V., Terrade, A., Krasny-Pacini, A., Kuppelin, M., Zinetti Bertschy, A., and Bossicard Schneider, A. (2022). *Cahier de participation - Groupe de Régulation des Emotions (manuel GREMO)*.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating Expressions of Opinions and

Emotions in Language. *Language Resources and Evaluation*, 39(2–3):165–210.

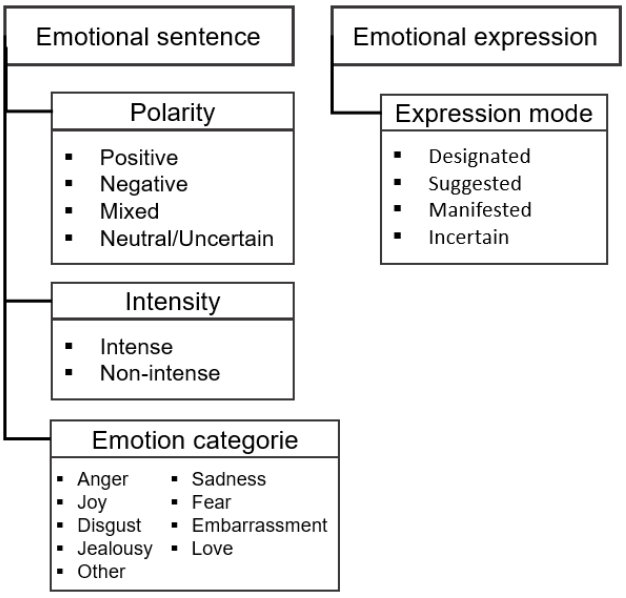
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., Zhang, M., Li, W., and Zhang, M. (2024). mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

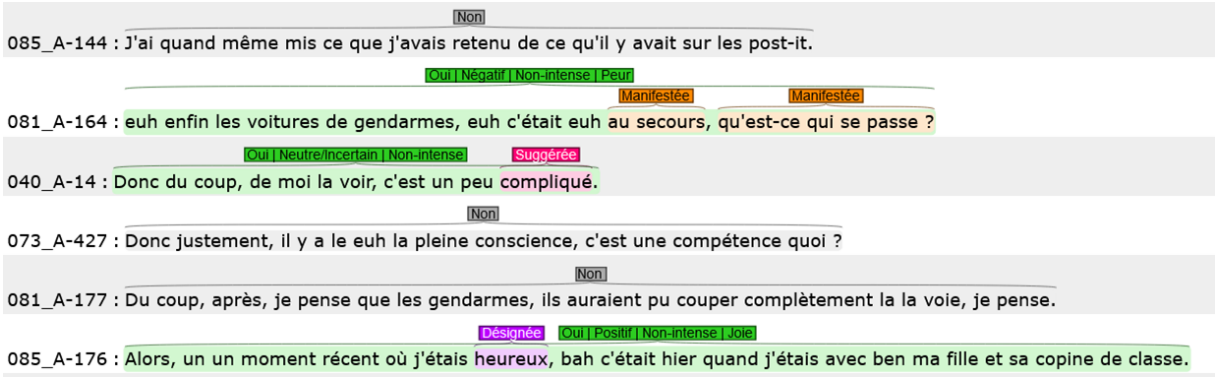
Appendix A. Detailed comparison with other works

Paper	Lg	Model	Corpus type	Emotional categories	Expression modes / emotion components
Öhman et al 2020	En	BERT	OPUS parallel movie subtitles (24,164 annotations)	Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust	
Casel et al. 2021	En	feature-based and deep-learning based approaches	Subcorpus of the REMAN (literature) & TEC (tweets) corpora (3041 instances)	Anger, Anticipation, Disgust, Fear, Joy, Neutral, Other, Sadness, Surprise, Trust	CPM components : Cognitive, Physiological, Action tendencies, Motor expressions, Subjective feelings
Cortal et al. 2023	Fr	DistillCamemBERT	Guided Narratives (3082 questionnaires answers)	Anger, Fear, Sadness, Joy	Behavior, Feeling, Thinking, Territory
Etienne et al. 2024	Fr	CamemBERT	Texts for children (5374 annotated sentences)	Admiration, Other, Anger, Guilt, Disgust, Embarrassment, Pride, Jealousy, Joy, Fear, Surprise, Sadness	Behavioral, Labeled, Displayed, Suggested
Noblet 2025	Fr	GTE	Opinion posts on innovations (4980 sentences)		CPM's evaluative component split in 4 dimensions : Familiarity, Pleasantness, Utility, Legitimacy
Ours	Fr	CamemBERTaV2	Transcriptions of patients discourse (7471 sentences)	Anger, Embarrassment, Disgust, Jealousy, Joy, Fear, Sadness, Love	Designated, Suggested, Manifested

Appendix B. Annotation scheme



Appendix C. Sentence- and token-level annotation in the INCEpTION annotation plateform



Appendix D. Prompts for annotation generation

1. Prompt for binary emotionality annotation:

prompt_emo = "Tu joues le rôle d'un expert linguiste qui annote des phrases en t'intéressant à leur expression émotionnelle.\

Définition : une phrase est dite 'émotionnelle' si elle exprime explicitement ou implicitement une émotion, qu'elle soit exprimée par le narrateur ou une autre personne.\

Question : La phrase à annoter est-elle **émotionnelle** ? Répond uniquement en utilisant les formes \"Oui\" ou \"Non\"\

Phrase à annoter: {phrase} \ "

2. Prompt for multilabel emotional category annotation:

prompt_categ_emo = "Tu joues le rôle d'un expert linguiste qui annote des phrases en t'intéressant à leur expression émotionnelle.

Si la phrase est émotionnelle, annote la catégorie émotionnelle précise de la phrase uniquement parmis les 8 catégories suivantes : [Amour, Colère, Dégoût, Honte, Jalousie, Joie, Peur, Tristesse].

Tu peux également dire que la phrase n'est pas émotionnelle. Dans ce cas, annote 'Non' et rien d'autre.

Tu ne peux choisir qu'une seule réponse. Donne moi juste le nom de la catégorie sans écrire de phrase.

Phrase à annoter: {phrase}”

3. Prompt for multilabel emotional expression mode annotation:

prompt_mode_expr = """Tu joues le rôle d'un expert linguiste qui annote des phrases en t'intéressant à leur expression émotionnelle.

Si la phrase est émotionnelle, le mode d'expression émotionnel utilisé dans la phrase uniquement parmis les 3 catégories suivantes : [Désignée, Suggérée, Manifestée].

L'émotion est appelée Désignée quand un mot ou syntagme en particulier renvoie directement à l'émotion en question et ne nécessite pas d'interprétation.

L'émotion est sémantiquement codée dans le mot et n\'a pas besoin d'être interprété.

Voici des exemples : « je suis épuisé » où « épuisé » renvoie au sentiment sans nécessiter d'interprétation ou encore « elle se sentait triste » où le mot triste évoque clairement le sentiment éprouvé.

Le terme doit pouvoir se définir avec les concepts suivants : états affectif, émotion, sentiment de, trait de personnalité.

S'il ne peut se définir par l'un de ces termes, alors il n'est pas à annoter en émotion explicite.

L'émotion est appelée Suggérée quand elle correspond à des situations, des objets, des interactions qui peuvent être responsables d'une émotion, des causes ou plutôt des raisons de son émergence.

Ce peut être un lien entre la situation et l'émotion basé sur des normes socio-culturelles : par exemple un anniversaire sera associé à la joie, un enterrement à la tristesse etc. Pour se rassurer qu'il s'agit d'une émotion suggérée, tu peux utiliser le schéma « Quand je suis dans la situation X, je me sens X ». Quelques exemples peuvent être : c'est nul (donc je n'aime pas, polarité=négative, classe=tristesse ou colère) etc.

L'émotion est appelée Manifestée quand elle passe par des traits extérieurs qui découlent logiquement d'un état intérieur. Par ex. je pleure parce que je me sens = triste, je ris parce que je me sens = joie.

Autrement dit, ce qui est exprimé dans la phrase est une conséquence d'une émotion sous-jacente.

Tu peux également dire que la phrase n'est pas émotionnelle. Dans ce cas, répond uniquement le mot 'Non' et rien d'autre.

Tu ne peux choisir qu'une seule réponse. Donne moi juste le nom de la catégorie sans écrire de phrase.

Phrase à annoter : {phrase} """