# Automatic Identification of Emotions in Texts: Dealing also with their Indirect Modes of Expression

**Anonymous ACL submission**

## Abstract

This paper presents a model that predicts whether (A) a sentence contains an emotion or not, (B) according to which mode(s) it is expressed, (C) whether the emotion is basic or complex, and (D) which emotional category it is. The originality of the paper lies in the focus on written texts (encyclopedia, novels, newspapers)—as opposed to the more widely studied conversational (sometimes multi-modal) situation—towards the analysis of text complexity in which emotions are one of the analysis factors according to certain works in psycho-linguistics. Within this particular scope, the major contribution of the paper is to introduce the identification of the modes of expression of the emotions, ranging from a direct lexical mode to the most indirect one where emotions are only suggested. The experiments are carried out on French texts for children. They show that the task is rather difficult but leading to acceptable results in comparison to what human annotators agree on. The results also seem to indicate that the task cannot be simply solved by prompting a large language model and requires a specialized model.

## 1 Introduction

When studying emotions in language, there are two main underlying topics: the categories of emotions which are reflected (i.e. their nature, their number) and the ways they are expressed (i.e. by which linguistic cues when analysing written texts, as it is our case study). In linguistics, emotions are recognized as a complex phenomenon, especially due to the diversity of linguistic markers used to express them, directly (via emotions names or labels, e.g., "happy" for category of joy, "regret" for category of guilt) or indirectly (e.g. description of events which are associated to emotional feelings mainly through social norms and conventions).This linguistic descriptive complexity obviously has its corollary when investigating processes involved in understanding of texts as it is pointed out in psycho-linguistics. NLP works on emotions usually only focus on one mode of expression of emotions: the emotions names or labels (e.g., "happy" for category of joy, "regret" for category of guilt). This limitation makes it difficult to integrate the emotional dimension for tasks that require fine-grained analysis and a better analysis of emotional density of texts, such as analyzing the complexity of texts from the point of view of different kinds of readers.

To this end, this papers proposes a new and original emotion identification model that introduces the notion of mode of expression in addition to the usual information on emotional categories (e.g., joy, fear, etc.). In practice, the model classifies the emotions in texts through four tasks: (A) predicting whether or not a sentence contains an emotion; (B) if so, how it is expressed (the *mode*); (C) whether it is a basic or complex category of emotion; and (D) which emotional category it falls into. This relies on the psycho-linguistically motivated annotation scheme proposed by Etienne et al. (2022), along with an annotated corpus of French texts. After adapting the data to the framed tasks, the model is instantiated as a multi-task CamemBERT model.

Evaluation shows that the proposed model outperforms expert approaches, non-neural approaches (SVM and XGBoost), and even ChatGPT. These conclusions are particularly interesting since, on the one hand, the proposed model significantly outperforms traditional approaches while remaining computationally affordable, and on the other hand, it tackles a problem for which a generic large language model like ChatGPT appears to be struggling. Moreover, the conducted human evaluation shows that the prediction errors made by the proposed model usually range in the same proportions as those made by humans.

In the remainder, Section 2 presents an overview of theoretical frameworks and work on emotion identification in texts in NLP. Section 3 formalises

and details the classification tasks A-D. Section 4 then details the experimental data, and preliminary choices on the final training process. Finally, Section 5 reports objective and human evaluations.

## 2 Literature Review

First of all, emotion analysis covers a significant portion of research focuses on *conversations* (Poria et al., 2019). This typically includes identifying when and how emotions are realized in chats/forums (Demszky et al., 2020a), speech transcripts (Zhou and Choi, 2018), or multi-modal dialogues (Busso et al., 2008; Poria et al., 2018; Chen et al., 2018). The present work does not belong to this area of research. Instead, it focuses on the problem of spotting and characterizing emotions in written texts ("in texts" for short), like in novels or articles.

Analyzing emotions in texts puts forward different aspects of emotions. As pointed out in (Klinger, 2023) and in (Troiano, 2023), more recent approaches of emotion analysis in NLP aim to gain a deeper understanding of which textual units support emotion predictions outside of emotional direct lexicon terms (e.g.,"happy", "anger"). To this end, they are more willing to rely on psychological and/or linguistic models of emotions. We adopt the same mindset here as we are interested at dealing both with direct and indirect modes of expression of emotions in texts. Like Troiano (2023), we aim to see how well computational models can be expected to perform when interpreting the kind of indirect expressed emotions. This section provides a brief overview of different approaches to studying emotions in texts and justifies the choice of the scheme and the data (annotated with this scheme) used in the rest of the paper. It also positions our work among the NLP studies on automatic emotion identification.

### 2.1 Frames to Study Emotions in Texts

A whole range of works in psycho-linguistics have studied the impact of characters' emotions on text comprehension, and have thus long shown the key role they play in the understanding process of texts (e.g., Dijkstra et al., 1995; Dyer, 1983). For instance, recent works highlighted two factors that influence children's understanding of emotions, and so of texts themselves: the *type of emotion* expressed, basic or complex —complex emotions (e.g.,pride, shame) being harder to grasp since they require knowledge on social norms— (Davidson, 2006; Blanc and Quenette, 2017); as well as *the way emotions are expressed* (Creissen and Blanc, 2017)), directly through an emotional label, indirectly through the mentioning of an emotional behaviour, or through the description of an emotional situation, the latter being the hardest to understand. Of course, the notion of emotional category is also addressed in psycho-linguistics, and it has been shown that some emotional categories take longer to be mastered by children (e.g., Baron-Cohen et al., 2010).

Many literature reviews in NLP (see for example Bostan and Klinger, 2018; Acheampong et al., 2020; Öhman, 2020) underline the great heterogeneity of emotion annotation schemes —and annotated corpora—, thus clearly highlighting the difficulty of emotion analysis. Indeed, this diversity affects all aspects of these works, from the notions (e.g., number and types of emotional categories) and type of data studied (e.g., newspapers, tweets), to the annotation procedures (e.g., crowd-sourcing, expert-annotation) and evaluation methods implemented (e.g., with or without inter-annotator agreement). Though some works endeavor to take into account broader sets of notions and linguistic cues to analyze emotions (e.g.,Casel et al., 2021; Kim and Klinger, 2019), the most commonly used concept remains the notion of *emotional category*, often tackled through a list of basic emotions introduced either by Ekman (1992) (anger, disgust, fear, joy, sadness, and surprise) or Plutchik (1980) (Ekman's categories, anticipation, and trust), with a focus on one way of expressing emotions: emotional lexicon.

The current work adopts the frame proposed by Etienne et al. (2022). It relies on a fine-grained emotion annotation scheme, and comes with a manually annotated corpus whose size is compatible with machine learning experiments. To our knowledge, this is the only work that could pertain to the goal of analyzing emotions in texts with dealing both direct and indirect ways of expressing them. It thus permits to access at a large coverage of emotions which is a key step for NLP .

### 2.2 Automatic Identification of Emotions

In NLP, emotion analysis in texts is usually seen as a classification task that requires emotion annotated corpora to develop models able to solve this task. As opposed to analysing conversations where several benchmarks exist, the heterogene-

ity of annotation schemes and annotated corpora (*cf.* Section 2.1) is then reflected in the diversity of classes predicted, items classified, and methods used to develop and evaluate the classifiers. Hence, the way results are presented also varies from one paper to another, which makes comparing performances harder. The following presents the main approaches and key trends in the results obtained.

**Predicted information.** While a few works only predict the presence/absence of emotional information in a given item (Alm et al., 2005; Aman and Szpakowicz, 2007), most propositions classify items according to the emotional categories. The focus is often on basic emotions (Strapparava and Mihalcea, 2007; Mohammad, 2012; Abdaoui et al., 2017; Demszky et al., 2020b; Öhman et al., 2020; Bianchi et al., 2021), though some works use a mix of basic and complex emotions (Balahur et al., 2012; Fraisse and Paroubek, 2015; Abdaoui et al., 2017; Mohammad et al., 2018; Liu et al., 2019; Demszky et al., 2020b). Furthermore, there is a long history of building and relying on emotional lexicons and the diversity of linguistic markers of emotions is not systematically taken into account, though mentioned in several works (Alm et al., 2005; Mohammad, 2012; Kim and Klinger, 2018; Demszky et al., 2020b)). Some works do study other means of expression. For instance, Kim and Klinger (2019) analyzes non-verbal expressions of emotions by characters in a corpus of fan fictions (e.g.,looks, gestures). Balahur et al. (2012) aim at detecting indirect emotions—that is to say, emotions not denoted by an emotional term—in a corpus of short texts describing situations in which the writer has felt an emotion. However, each these works focus on a unique way of expressing emotions (non-verbal expressions of emotion in (Kim and Klinger, 2019) and indirect emotions in (Balahur et al., 2012)). For their part, based on Scherer's (2005) emotion component process model, (Casel et al., 2021) annotated and then predicted several components of emotions, such as physiological symptoms and motor expressions of emotions, or cognitive appraisal of events. Though (Casel et al., 2021) deals with a broader set of cues, those are not strictly linguistically motivated. Hence, by relying on (Etienne et al., 2022), the true originality of our work lies in the consideration of different modes of expression of emotions not solely direct ones (see details in Section 3.2).

**Technical approaches.** Historically, *Support Vector Machine* (SVM) models have been widely used to classify sentences (Aman and Szpakowicz, 2007; Mohammad, 2012)) or texts (Abdaoui et al., 2017; Balahur et al., 2012; Fraisse and Paroubek, 2015; Mohammad, 2012) according to the emotional category they express. Until the rise of the embeddings, inputs were mostly symbolic: bags-of-words or *n*-grams, features based on emotional resources such as WordNetAffect (Aman and Szpakowicz, 2007; Balahur et al., 2012; Strapparava and Mihalcea, 2007) or emotional lexicons (Strapparava and Mihalcea, 2007; Abdaoui et al., 2017; Kim and Klinger, 2018).

More recently, neural networks (Kim and Klinger, 2018) and Transformer architectures (Liu et al., 2019; Demszky et al., 2020b; Öhman et al., 2020; Bianchi et al., 2021) trust the state of the art, although performances are uneven from class to class, with F1 scores rarely above 0.75-0.80. Regarding French, no Transformer-based model has been proposed yet to our knowledge.

Overall, the NLP literature shows that the automatic emotion analysis is a complex NLP task, that even Transformer models do not manage to solve entirely. To monitor the progress of this task, experiments will compare the results of our proposed model with related work.

## 3 Tasks and Proposed Model

Constructed in the global perspective of analyzing the complexity of texts, the aim of this work is to propose a Transformer model for identifying emotions that integrates 4 different levels of emotion analysis. Each level corresponds to a machine learning task treated as a classification problem. The granularity used for prediction is the *sentence level* The resolution of these tasks (referred to as Tasks A to D) is carried at the sentence level, as opposed to the text level. This enables, for instance, studying how the presence of emotions can evolve along a text. It is important to highlight that sentences can contain several emotions. Hence, all tasks are *multi-label* classification tasks. Furthermore, all tasks are learned together in a multi-task fashion, leading to a unique model. This section first details each task, and then the models' settings.

### 3.1 Task A: Presence of Emotion

The first task aims at predicting the presence of emotional information in a given sentence (binary prediction). It offers two advantages: on the one hand, it constitutes the easiest task in automatic

emotion analysis; on the other hand, it reflects a text complexity factor. Indeed, it has been shown that the mere presence of emotional information (no matter the emotional category expressed or how it is expressed) can enhance text comprehension (e.g., for children in Davidson et al., 2001).

### 3.2 Task B: Expression Mode

As mentioned in Section 2, the expression mode focuses on the linguistic means used to convey the presence of an emotion in a text. It allows for a finer and broader linguistic analysis of emotion, and it also reflects a complexity marker of texts. Following Etienne et al. (2022), Task B considers 4 expression modes. The first one is a direct mode while the next three ones can be seen as indirect modes: **Labeled emotions** which refer to emotions directly denoted by an emotional label, *i.e.* an emotional lexicon term (e.g., *happy*, *scare*); **Behavioral emotions** which refer to emotions expressed by the description of an emotional behaviour, such as physiological manifestations (e.g., *cry*, *smile*), or more complex behaviours (e.g., *to slap someone*); **Displayed emotions** which refer to emotions expressed by very heterogeneous surface linguistic characteristics of utterances that reflect mainly the speaker's emotional state (e.g., interjections, short sentences); **Suggested emotions** which refer to emotions expressed by the description of a situation associated to an emotional feeling through social norms and conventions (e.g., *seeing a good friend after a long time* suggests joy).

### 3.3 Task C: Emotion Type

Task C aims at predicting the presence of two emotion types (*basic* and *complex*) in a given sentence (2 simultaneous binary predictions). To our knowledge, this notion has yet never been studied as is in automatic emotion analysis (even though both *basic* and *complex* emotional categories have been used in NLP (*cf.* Section 2.2)). This is likely due to the fact that the type of an emotion expressed is directly linked to its emotional category. However, the notion of emotion type constitutes a complexity marker in itself, since complex emotions are harder to understand for children for example (*cf.* Section 2.1). That is why we decided to design a task specifically aimed at identifying this notion.

### 3.4 Task D: Emotional Category

Most widely used in NLP emotion analysis tasks, the notion of emotional category also constitutes a complexity marker in text for children (*cf.* Section 2.1). In line with Etienne et al. (2022), Task D is designed to label 11+1 emotional categories, namely Ekman's 6 basic emotions (*anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*) and 5 complex emotions (*admiration*, *embarrassment*, *guilt*, *jealousy*, and *pride*). A last category, named *other*, is used to capture markers that express any other emotion (e.g., hatred, disdain, love, *etc.*).

### 3.5 Proposed Model

The proposed model results from fine-tuning the base version of the pre-trained transformer model CamemBERT (Martin et al., 2020)[1]. This is 110M-parameters 12-layers encoder model based on BERT and trained on 138GB of French texts (Suárez et al., 2019). While fine-tuning more recent and larger (generative) language models like Llama2 or Mistral would probably lead to better results, the choice of a reasonably-sized model is motivated by two reasons. First, the objective of the paper is to show that, contrary to several other NLP tasks, fine-grained emotion characterization in texts cannot be performed by prompting generic (i.e., non-specialized) large language models. The second reason is that our work target a light-weight solution, such that emotion characterization can be integrated as a processor for text complexity analysis in a massive collection of texts of a public search engine. So, even if providing results for larger fine-tuned models is in the roadmap, this is not performed in this paper.

In practice, the baseline CamemBERT model's fine-tuning is updated by replacing its last token prediction layer with a binary classification layer, and no layer is frozen. To bootstrap the model, a first fine-tuning was performed on the sole binary Task A for 3 epochs, since this seemed to be the easiest task to start with. The final multi-task model is a continuation of this first fine-tuning during 6 more epochs where the classification layer is extended to integrate Tasks B, C, and D.[2] Given an input sample, the final model handles all tasks (A-D) at once.

## 4 Data and Developmental Work

This section first describes the data used to train the model before commenting on preliminary experiments conducted to train this model.

[1] https://huggingface.co/camembert-base

[2] For each fine-tuning, the optimizer is Adam with a learning rate of 10-5 (w/o decay) and a batch size of 8.

### 4.1 Data

The corpus is the one provided along the annotation scheme (Etienne et al., 2022). It is made of 1,594 French texts (28K sentences) destined to children from 6 to 14 years old, dispatched into 3 genres: newspaper. It comes with expert annotations delimiting emotional units (chunks) within the texts. Each unit is described with its expression mode and emotional category.

**Preparation** Annotations at the chunk level were merged to the sentence level. Hence, a given sentence can cover several emotional units (hence the consideration of multi-label classifications). The presence of emotion and emotion types were derived from the expression mode and emotion category labels. In the end, each sentence is associated with a vector of 19 flags. Examples of (in-context) sentences are provided in Table 1.The data is split into train, dev, and test subsets based on the number of sentences (70/10/20%, respectively). Partitioning is such that all sentences of a text are situated in the same subset, in order to avoid training bias based on particularities of texts (e.g., the name of a character).

**Distribution** Table 2 presents the proportion of labels of each classification task within the subsets of the corpus. Several imbalances emerge. (A) Only 15-20% of sentences are emotional, with percentages similar across subsets. This observation applies to all other labels. (B) Expression modes all represent similar proportions, though the label 'displayed' is the less frequent (3% of the sentences) and 'suggested' the most common (6%). When combined, the percentages of expression modes labels are higher than those of the 'emotional' label. This is due to the fact that a sentence can hold several emotional units, expressed by different expression modes. (C) Emotion types labels are greatly imbalanced, with a clear dominance of basic emotions. The sum of 'basic' and 'complex' labels percentages is lower than those of emotional sentences, most likely because emotional units expressing the emotional category 'other' are not associated to an emotion type. (D) Like emotion types labels, emotional categories labels are highly imbalanced. Dominant labels are 'anger', 'fear', 'joy', 'sadness', 'surprise', and 'other'. Emotional categories labels are scarce in general (always under 5% of sentences), but some are even rarer (though still present), in particular 'disgust', 'guilt', and 'jealousy'.

### 4.2 Developmental Work

To obtain the best classifiers, several factors influencing their performances have been experimented on the development set.

**Corpus balancing** Different strategies have been studied to tackle the imbalance of the classes (see Section 4.1), including subsampling majority classes and weighting classification losses according to the inverse frequency of each class. No significant difference in performances was observed and no strategy is used in the final experiments.

**Contextualization** Intuitively, inferring an emotion from a sentence requires to understand the context. In this regard, annotations from (Etienne et al., 2022) were carried out based on the entire text. Hence, we compared training the model either based on single (context-free) sentence, or triplets of sentences where the target sentence to label is surrounded by its previous and following sentence (if they exist)[3]. For all tasks, better performances were obtained with contextualization. This is what is used in the final experiments.

## 5 Automatic and Human Evaluations

This section compares our model with other approaches using automatic evaluation, and provides a qualitative analysis through human evaluation.

### 5.1 Comparison with Baselines

To better grasp the feasibility of Tasks A-D and the suitability of CamemBERT to solve them, the proposed model is compared to 3 types of baselines.

**SVM** SVM models were trained since this is an historical approach in the field (see Section 2.2). Two types of input features were used: (i) Bag-of-tokens where tokens are from CamemBERT's tokenizer, restricted to those observed on the training set - the resulting size of the input vector is 18,437; (ii) Sentence embeddings of size 768 obtained with SentenceTransformer (Reimers and Gurevych, 2019) and CamemBERT[4].

**XGBoost** XGBoost models were trained since this is a more recent competitive light-weight technique for many classification tasks, especially with unbalanced data (Chen and Guestrin, 2016). Settings for input features are the same as for SVMs.

**ChatGPT** Since many NLP tasks have recently been outperformed by large language models, our

[3]Format is "before: preceding</s>current: target</s>after: following</s>".

[4]https://huggingface.co/dangvantuan/sentence-camembert-base

| Sentence (+ surrounding sentences) | (A) | (B) Express. mode | | | | (C) Em. type | | (D) Emotion category | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | emo. | beha. | lab. | disp. | sug. | bas. | comp. | adm. | other | ang. | guilt | dis. | emb. | pride | jeal. | joy | fear | surp. | sad. |
| How does the coronavirus spread? **Especially through respiratory droplets expelled by an infected person.** Respiratory droplets are small droplets of saliva that are released into the air when we talk, cough, or sneeze. | | | | | | | | | | | | | | | | | | | |
| It is mainly celebrated in the Anglo-Saxon world. **Traditionally, children wear scary costumes.** They dress up as often despised and feared creatures such as ghosts, vampires, or witches and go door-to-door in the neighborhood, asking for candies or pastries. | ✓ | | ✓ | | | ✓ | | | | | | | | | | | ✓ | | |
| – He succumbed after ingesting his herbal tea and a toxic substance, presumably cyanide. **From there, it was only a small step for Angus's mother to accuse the king of murder as she rushed towards her brother.** – The herbal tea... | ✓ | ✓ | | | ✓ | ✓ | | | | ✓ | | | | | | | | | |
| This summer, Nolita had to eat a sausage for the first time in a long time because there was nothing else. "I forced myself," she said. **"It disgusted me, and I felt guilty," she recounted.** | ✓ | | ✓ | | | ✓ | ✓ | | | | ✓ | ✓ | | | | | | | |
| At the Rome Olympics, the historic event takes place during the marathon: Ethiopian Abebe Bikila becomes the first athlete from black Africa to become an Olympic champion. **What's more, he achieved this feat... barefoot!** He had indeed developed the habit of running barefoot back home in Ethiopia. | ✓ | | | ✓ | ✓ | ✓ | ✓ | | | | | | | ✓ | | ✓ | | ✓ | |

Table 1: Examples of sentences (translated from French) and ref. labels for Tasks A, B, C, and D.

| Task | Labels | Sent. Prop. (%) | | |
|---|---|---|---|---|
| | | train | dev | test |
| (A) Pres. of emotion | emotional | 20.2 | 15.8 | 17.6 |
| (B) Expression mode | behavioral | 4.6 | 3.6 | 4.3 |
| | labeled | 5.3 | 5.2 | 5.7 |
| | displayed | 3.6 | 2.3 | 3.5 |
| | suggested | 7.1 | 5.8 | 6.3 |
| (C) Emotion type | basic | 15.4 | 12.6 | 13.9 |
| | complex | 2.0 | 2.1 | 2.3 |
| (D) Emotional category | admiration | 0.6 | 1.1 | 1.0 |
| | other | 5.0 | 3.2 | 3.7 |
| | anger | 4.6 | 3.2 | 3.4 |
| | guilt | 0.1 | 0.0 | 0.1 |
| | disgust | 0.2 | 0.3 | 0.2 |
| | embarrass. | 0.6 | 0.6 | 0.6 |
| | pride | 0.7 | 0.4 | 0.9 |
| | jealousy | 0.0 | 0.0 | 0.0 |
| | joy | 3.2 | 2.3 | 3.6 |
| | fear | 3.8 | 3.3 | 3.8 |
| | surprise | 3.0 | 3.1 | 2.5 |
| | sadness | 2.5 | 2.0 | 2.5 |

Table 2: Distribution of labels

| Task | Model | Macro R | Macro P | Macro F1 |
|---|---|---|---|---|
| (A) Presence of emotion | SVM | 0.481 | 0.659 | 0.556 |
| | XGBoost | 0.223 | 0.700 | 0.338 |
| | ChatGPT | 0.622 | 0.443 | 0.518 |
| | ours | **0.764** | **0.741** | **0.752** |
| (B) Expression mode | SVM | 0.267 | 0.721 | 0.368 |
| | XGBoost | 0.218 | **0.730** | 0.314 |
| | ChatGPT | 0.513 | 0.101 | 0.152 |
| | ours | **0.626** | 0.665 | **0.645** |
| (C) Emotion type | SVM | 0.211 | 0.343 | **0.261** |
| | XGBoost | 0.120 | **0.659** | 0.200 |
| | ChatGPT | **0.756** | 0.123 | 0.199 |
| | ours | 0.557 | 0.662 | **0.601** |
| (D) Emotional category | SVM | 0.125 | 0.487 | 0.186 |
| | XGBoost | 0.192 | **0.565** | 0.272 |
| | ChatGPT | **0.697** | 0.109 | 0.174 |
| | ours | 0.397 | 0.463 | **0.420** |

Table 3: Performances of baseline models

approach is compared to ChatGPT (Ouyang et al., 2022). For a given input sample, ChatGPT is prompted in a few-shot manner to annotate it with binary labels (yes/no). Consecutively for each task and label, a natural language description of what is expected is given to the model before asking to answer. Examples from the training set are also provided for each label.[5] The approach was tested with different prompts, including either (i) 2 up to 4 positive examples only or (ii) 3 or 4 negative examples as well. For one input sample to label for all tasks, the prompts are 4,000 and 6,000 tokens long, for cases (i) and (ii), respectively. Hence, contrary to SVM and XGBoost, the approach is not frugal but it does not require any training.

Table 3 sums up performances of best baseline models for each task and compares them to our proposed model. For SVMs and XGBoost, these performances are reached with the bag-of-tokens inputs. For ChatGPT, the best prompts are the ones without negative examples.[6] Models are evaluated through recall (R), precision (P), and F-measure (F1) scores. When given at task level (as opposed to label level), performances correspond to macro-averages, i.e., the same weight is given to each label predicted. For each approach, variants have been experimented on the development set and all results presented here are obtained on the test set. Overall, it appears that our proposed model significantly outperform SVM, XGBoost and ChatGPT regarding F1 scores for all tasks, with values which are almost double of the second best ranked model for Tasks B, C and D. It especially appears that all baselines tend to favor either recall (ChatGPT) or precision (SVM, XGBoost), whereas our model is balanced. Finally, results of ChatGPT are not low. This shows that the task is difficult and requires a specific expertise or training.

[5]To make it unambiguous, examples are always mono-label for the target task under consideration.

[6]If accepted, all prompts and results will be in appendices.

| Ref. | Lg. | Labels | Model | Lexicon | Granularity | Macro-F1 of best model |
|---|---|---|---|---|---|---|
| ours | FR | anger, disg., joy, fear, surpr., sadn. | Transformer | none | sent. triplets | 0.52 |
| (Öhman et al., 2020) | EN | same + trust, anticipation | Transformer | none | sent. | 0.54 |
| (Kim and Klinger, 2018) | EN | same + trust, anticipation | symb. | NRC lexicon | sent. triplets | 0.31 |
| | | | MLP | none | sent. triplets | 0.31 |
| (Fraisse, Paroubek, 2015) | FR | anger, fear, sadness | SVM | custom | paragr. | 0.31 |

Table 4: Comparison with diverse works.

| Task | Label | Approach | Macro-F1 |
|---|---|---|---|
| (A) Pres. of emotion | emotional | ours | **0.752** |
| | | TextBlob | 0.299 |
| | | Emotaix | 0.445 |
| (B) Expression mode | behavioral | ours | **0.626** |
| | | Emotaix | 0.041 |
| | labeled | ours | **0.807** |
| | | Emotaix | 0.559 |
| (D) Emot. categories (labeled mode only) | all | ours | **0.466** |
| | | Emotaix | 0.425 |
| Emotion polarity | positive | ours | **0.575** |
| | | TextBlob | 0.163 |
| | negative | ours | **0.678** |
| | | TextBlob | 0.168 |

Table 5: Comparison on tasks that can be derived from currently available public resources.

### 5.2 Comparison with Related Work

Our classifier's performances cannot be directly compared with those of other NLP emotion analysis models, due to the lack of any work truly similar. As we said before, the very major difference is the inclusion of the expression modes.Nonetheless, this section reports complementary results to provide a better intuition of how our model performs.

**Closest comparable works** Table 4 sums up performances from the 3 closest works we could find from the scientific literature. These works have been chosen because they all predict labels at the same (or at a similar) granularity level as our own classifier. More precisely: (Öhman et al., 2020) allows for a comparison with another Tansformer model; (Fraisse and Paroubek, 2015) with another work on French; and (Kim and Klinger, 2018) with another NLP work which uses emotion annotation at the linguistic marker level (*v.s.* at a sentence or text level). All of them only focus on the sole emotional categories. What emerges from this table is that, whatever model is put in comparison, our classifier always performs as well or even better.

**Existing resources** Two resources available for French are interesting for emotion identification: TextBlob[7], a sentiment analysis library that embeds a French lexicon where terms are associated with a negative and positive weight reflecting their polarity; Emotaix (Piolat and Bannour, 2009), another lexicon where terms are associated to emotional categories for the sole labeled mode. Emotaix also provides terms associated with the behavioral modes, but here without any information about the corresponding emotional category. Several tasks handled by our model were replicated with TextBlob and Emotaix. To take into account scope differences between these resources and our proposed model, Task B was limited to the sole behavioral and labeled modes while Task D was limited to the labeled mode. Furthermore, our model got experimented on an emotion polarity prediction task on our test set since TextBlob is designed for this use case. To predict the polarity based on our model, emotional categories were predicted, and empirically mapped to the positive or negative polarity (e.g.,'anger' is 'negative', 'joy' is 'positive'). Table 5 reports results obtained by implementing classifiers based on these resources. Overall, they show that the proposed model performs significantly better than TextBlob and Emotaix, including in the emotion polarity task for which it was not specifically designed. The only task for which competition remains is the prediction of categories when the mode is labeled, which is the easiest situation compared to considering all modes.

### 5.3 Detailed Analysis

This section details the results for each task, and then presents the human evaluation.

**Label-wise results** Table 6 presents detailed results of our classifier on the prediction of the 19 labels of all tasks. Additional observations that can be made are as follows. (B) Regarding the expression modes, it can be noted that labeled emotions are very well recognized (F1 > 0.8), as opposed to suggested emotions (F1 < 0.5). This comes as no surprise, since labeled emotions are the easiest for human annotators to identify, while suggested emotions come with the blurriest boundaries, and are therefore the most difficult for humans themselves to analyze. (C) The prediction performance for emotion types appears for its part to be linked to the results of the prediction of emotional categories, as the label 'basic' is better recognized than the 'complex' label. (D) Finally, 3 emotional

[7] https://textblob.readthedocs.io/

| Task | Macro R | Macro P | Macro F1 | Labels | R | P | F1 |
|---|---|---|---|---|---|---|---|
| (A) Pres. of emotion | 0.764 | 0.741 | 0.752 | emo. | 0.764 | 0.741 | 0.752 |
| (B) Express. mode | 0.626 | 0.665 | 0.645 | beha. | 0.601 | 0.653 | 0.626 |
| | | | | lab. | 0.811 | 0.803 | 0.807 |
| | | | | disp. | 0.667 | 0.726 | 0.695 |
| | | | | sug. | 0.426 | 0.479 | 0.451 |
| (C) Emot. type | 0.557 | 0.662 | 0.601 | bas. | 0.705 | 0.733 | 0.719 |
| | | | | comp. | 0.409 | 0.591 | 0.484 |
| (D) Emot. category | 0.397 | 0.463 | 0.420 | adm. | 0.281 | 0.457 | 0.348 |
| | | | | other | 0.745 | 0.592 | 0.660 |
| | | | | ang. | 0.670 | 0.685 | 0.677 |
| | | | | guilt | 0.000 | 0.000 | 0.000 |
| | | | | dis. | 0.000 | 0.000 | 0.000 |
| | | | | emb. | 0.364 | 0.600 | 0.453 |
| | | | | pride | 0.333 | 0.615 | 0.432 |
| | | | | jeal. | 0.000 | 0.000 | 0.000 |
| | | | | joy | 0.530 | 0.709 | 0.606 |
| | | | | fear | 0.717 | 0.661 | 0.688 |
| | | | | surp. | 0.697 | 0.739 | 0.717 |
| | | | | sad. | 0.428 | 0.504 | 0.463 |

Table 6: Performances of best model

| Source of label | Evaluator's opinion | Proportion (num. of labels) emot. category | expr. mode |
|---|---|---|---|
| human & model | Agree | 95.5% (105) | 97.7% (129) |
| | Disagree | 4.5% (5) | 2.3% (3) |
| model | Agree | 92.1% (58) | 91.1% (41) |
| | Disagree | 7.9% (5) | 8.9% (4) |
| model | Agree | 76.5% (39) | 90.2% (37) |
| | Disagree | 23.5% (12) | 9.8% (4) |

Table 7: Agreement of experts.

categories are never predicted ('guilt', 'disgust', and 'jealousy'). They are the rarest labels of the training set and there was therefore probably not enough occurrences for the model to learn how to predict them. As a matter of fact, best predicted emotional categories are basic, more frequent emotions, namely labels 'surprise', 'fear', 'anger' (*cf.* Table 2). However, even though 'surprise' is the best predicted label of task D, it is not the most represented in the training set. On the contrary, 'sadness' is not well recognized, even though it is one of the most frequent emotional categories. This can be explained by the interaction between ¨the expression mode and the emotional category. Indeed, if there is a strong association between an emotional category and an expression mode in the training set, the model will recognize the category better when it is expressed by this mode. For instance, complementary analysis shows that 'surprise', which is predominantly displayed in the corpus, is on average 14 times better recognized by the model when it is displayed than when it is expressed by other modes. 'Anger', which is mostly behavioral, is on average 4 times better recognized when expressed by this mode.

**Human evaluation** Given the difficulty of the considered tasks, it is important to cross-reference the automatic evaluation with human analysis, especially to provide an intuition of what the observed prediction errors represent. To investigate this aspect, a perceptual validation experiment was carried out with 3 experts in text complexity and emotions. Each of them was informed of the tasks and definition choices underlying each label in psycholinguistics and linguistics. They were then each confronted with 150 sentences from the test set and their associated emotional category and expression mode labels. These labels either came from the *human* annotations, or from our *model*'s predictions. For each label, the experts had to say whether they agreed or not with the proposed annotation.

Table 7 reports the percentages of experts' agreement[8] with the proposed labels, depending on the source of the label ("human" for reference annotations or "machine" for model's predictions). Though strongest agreement is when both the human and model's labels match ("human & model"), agreement scores are overall very high, especially for expression mode. These preliminary results thus show that, even when the model offers an analysis divergent from the reference, it is usually seen as relevant by human experts. This demonstrates that our model is able to generalize properly and that all F1 scores from previous experiments underestimate the model's perceptual quality.

## 6 Conclusion and Perspectives

We have proposed a model for analysing emotions which is original in NLP because it takes into account their direct but also their indirect modes of expression. Furthermore, experiments show that this model pushes performance on the studied tasks to a level that previous works or off-the-shelf solutions were not able to achieve until now. Human evaluation showed that this level is almost equivalent to what humans can do. The model and data will be made publicly available.

In the future, a straightforward application of our model is complexity analysis—the broader context in which this work has been carried out—, since the labels predicted reflect complexity markers (expression modes and emotion types in particular). More broadly, our work could contribute in psychological research to study the link between emotional language and the psychological state of the writer/speaker, along the same line as the studies reported in (Tausczik and Pennebaker, 2010).

[8]We considered that experts agree with a label when at least 2 out of 3 stated that they agree with the label.

## 7 Limitations

The expression of emotions is a complex phenomenon. In the specific context of texts, some notions that are absent from our work deserve to be taken into account in the future. Particularly, the notion of experiencer is important as it explains who experiences an emotion and provides an additional key in analyzing the complexity of a text (e.g.,cases where there are multiple characters or multiple emotions, legitimacy of an emotion in relation to a character's situation). Furthermore, the notion of experiencer allows for the use of theoretical frameworks for the relationships between characters, the narrator, and the reader. To explore the notion of experiencer, the present work would require modifying the NLP task into a more complex task of generating structured annotations of emotional units instead of binary classification at the sentence level.

Then, although the focus is on the (psycho-)linguistic expression of emotions in texts, our results would benefit from being compared with those from the community interested in analyzing emotions in conversational or even multimodal interaction settings. This community is more rooted in signal processing and machine learning. The advances from this community would likely contribute to improving our work in terms of the technical implementation of the model. However, this work has not been carried out here due to the communities' differences, especially in relation to the complexity analysis project in which our work is situated.

## 8 Ethics Statement

Given the sensitive nature of emotions and the public availability of our model, the use of the results it provides must be done in a responsible and ethical manner. It is crucial to consider the potential consequences of using these results, avoiding any abusive use on non-consented data or with the intention of manipulating/influencing emotions. Furthermore, emotions can be influenced by cultural or social biases. It is important to ensure, in the context of using our model, that these biases do not lead to discrimination or unfair prejudices in the obtained results.

## References


Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2017. Feel: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855. Publisher: Springer.

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189. Publisher: Wiley Online Library.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from Text: Machine Learning for Text-based Emotion Prediction. In *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.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer.

Alexandra Balahur, Jesús M Hermida, and Andrés Montoyo. 2012. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision support systems*, 53(4):742–753. Publisher: Elsevier.

Simon Baron-Cohen, Ofer Golan, Sally Wheelwright, and Yael Granader. 2010. Emotion Word Comprehension from 4 to 16 Years Old: A Developmental Survey. *Frontiers in Evolutionary Neuroscience*, 0. Publisher: Frontiers.

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

Nathalie Blanc and Guy Quenette. 2017. La production d'inférences émotionnelles entre 8 et 10 ans : quelle méthodologie pour quels résultats ? *Enfance*, 4(4):503–511. Place: Paris Publisher: NecPlus.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An Analysis of Annotated Corpora for Emotion Classification in Text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Felix Casel, Amelie Heindl, and Roman Klinger. 2021. Emotion recognition under consideration of the emotion component process model. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 49–61, Düsseldorf, Germany. KONVENS 2021 Organizers.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

S. Creissen and N. Blanc. 2017. Quelle représentation des différentes facettes de la dimension émotionnelle d'une histoire entre l'âge de 6 et 10ans ? apports d'une étude multimédia. *Psychologie Française*, 62(3):263–277. Cognition et multimédia : les atouts du numérique en situation d'apprentissage.

Denise Davidson. 2006. The Role of Basic, Self-Conscious and Self-Conscious Evaluative Emotions in Children's Memory and Understanding of Emotion. *Motivation and Emotion*, 30(3):232–242.

Denise Davidson, Zupei Luo, and Matthew J. Burden. 2001. Children's recall of emotional behaviours, emotional labels, and nonemotional behaviours: Does emotion enhance memory? *Cognition and Emotion*, 15(1):1–26. Place: United Kingdom Publisher: Taylor & Francis.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020a. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. 2020b. GoEmotions: A Dataset of Fine-Grained Emotions. *CoRR*, abs/2005.00547. ArXiv: 2005.00547.

Katinka Dijkstra, Rolf A Zwaan, Arthur C Graesser, and Joseph P Magliano. 1995. Character and reader emotions in literary texts. *Poetics*, 23(1-2):139–157. Publisher: Elsevier.

Michael G Dyer. 1983. The role of affect in narratives. *Cognitive science*, 7(3):211–242. Publisher: Wiley Online Library.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200. Publisher: Routledge _eprint: https://doi.org/10.1080/02699939208411068.

Aline Etienne, Delphine Battistelli, and Gwénolé Lecorvé. 2022. A (Psycho-)Linguistically Motivated Scheme for Annotating and Exploring Emotions in a Genre-Diverse Corpus. In *13th Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Amel Fraisse and Patrick Paroubek. 2015. Utiliser les interjections pour détecter les émotions. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 279–290.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359.

Evgeny Kim and Roman Klinger. 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.

Roman Klinger. 2023. Where are we in event-centric emotion analysis? bridging emotion role labeling and appraisal-based approaches.

Chen Liu, Muhammad Osama, and Anderson de Andrade. 2019. DENS: A Dataset for Multi-class Emotion Analysis. *CoRR*, abs/1910.11769. ArXiv: 1910.11769.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 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.

Saif Mohammad. 2012. #Emotional Tweets. In **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.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Annie Piolat and Rachid Bannour. 2009. An example of text analysis software (emotaix-tropes) use: The influence of anxiety on expressive writing. *Current psychology letters. Behaviour, brain & cognition*, 25(2, 2009).

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

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.

Klaus R Scherer. 2005. What are emotions? And how can they be measured? *Social science information*, 44(4):695–729. Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

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

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Oberländer L. Klinger R. Troiano, E. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1).

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Emily Öhman. 2020. Emotion annotation: Rethinking emotion categorization. *CEUR Workshop Proceedings*, 2865:134–144. Publisher: CEUR-WS.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.