Constant Bonard, Gustave Cortal

To cite this version:

HAL Id: hal-04624340
https://hal.science/hal-04624340v2
Submitted on 3 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Constant Bonard
University of Bern, Institute of Philosophy, Hochschulstrasse 4, 3012 Berne, Suisse
constant.bonard@gmail.com

Gustave Cortal
Université Paris-Saclay, ENS Paris-Saclay, CNRS, LMF, 91190, Gif-sur-Yvette, France
Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
gcortal@ens-paris-saclay.fr

Abstract

We propose leveraging cognitive science research on emotions and communication to improve language models for emotion analysis. First, we present the main emotion theories in psychology and cognitive science. Then, we introduce the main methods of emotion annotation in natural language processing and their connections to psychological theories. We also present the two main types of analyses of emotional communication in cognitive pragmatics. Finally, based on the cognitive science research presented, we propose directions for improving language models for emotion analysis. We suggest that these research efforts pave the way for constructing new annotation schemes, methods, and a possible benchmark for emotional understanding, considering different facets of human emotion and communication.

1 Introduction

Emotion analysis in natural language processing aims to develop computational models capable of discerning human emotions in text. Recently, language models have been widely used to solve various tasks in natural language processing, including emotion analysis (Devlin et al., 2019; Brown et al., 2020). This field of research faces several limitations. First, different ways of conceptualizing emotions lead to different annotation schemes and datasets (Klinger, 2023). As a result, the generalization ability of models is limited, and it is often impossible to compare studies. To address these limitations, it has been proposed to unify some annotation schemes based on the semantic proximity of emotion categories (Bostan and Klinger, 2018), to automatically find emotion categories from data (De Bruyne et al., 2020), or to obtain emotion embeddings independent of annotation schemes (Buechel et al., 2021). Inspired by psychology and cognitive science research, we believe building an annotation scheme unifying different perspectives on the emotional phenomenon would be possible and desirable.

In addition, existing benchmarks evaluate certain aspects of emotional understanding but do not consider its full complexity (Campagnano et al., 2022; Zhang et al., 2023a; Paech, 2024). For example, Paech (2024) proposes to evaluate the emotional understanding of language models by predicting the intensity of emotions in conflict scenes. This type of evaluation is too limited: benchmarks should reflect as much as possible the richness of emotional understanding in humans, a richness documented in different branches of affective sciences (Green, 2007; Wharton, 2016; Scarantino, 2017; Barrett et al., 2019; Bonard and Deonna, 2023).

Another related research area focuses on the theory of mind of language models, i.e., their ability to correctly attribute mental states to others. In our view, this literature is promising in that it links recent developments in language models to theories and empirical methods in cognitive science (for a review, see Bonard (2024, section 5)). Notably, several tasks and benchmarks have been developed to measure the ability of language models to succeed at different versions of the False Belief Task (Trott et al., 2022; Aru et al., 2023; Gandhi et al., 2023; Holerman and van Deemter, 2023; Kosinski, 2023; Mitchell and Krakauer, 2023; Shapira et al., 2023; Stojnić et al., 2023; Ullman, 2023). However, theory of mind and, more generally, social reasoning abilities go beyond the ability to succeed at the False Belief Task (Apperly and Butterfill, 2009; Langley et al., 2022; Ma et al., 2023). The ability to correctly interpret expressed emotions cannot be reduced to it. The degree to which language models possess this emotional competence is worth studying in its own right.

Generally speaking, research on language mod-
els for emotion analysis would benefit from cognitive science research on emotion and communication. In particular, we believe this approach can lead to better ways of annotating emotions expressed in text. Additionally, it can improve the evaluation of the emotional understanding of language models by developing new benchmarks. In what follows, we present an overview of psychological theories of emotion (section 2) and ways of annotating emotions in natural language processing (section 3). Then, inspired by specific psychological and linguistic theories (section 4), we propose research directions to address some of the current limitations of emotion analysis (section 5).

**Contributions.** We propose integrating different cognitive science theories on emotion with NLP research. We explain why and how emotion analysis should use research from cognitive pragmatics, specifically what we call “the detective analysis”, to improve automatic emotion analysis. We suggest that these points lead both to devising a new annotation scheme and improving how language models should be evaluated for emotion analysis.

### 2 Emotion Theories in Cognitive Science

This section will present the three main emotion theories in psychology to provide a background for connecting emotion analysis in natural language processing with cognitive science.

**Basic emotion theory.** Basic emotion theory is certainly the most influential today. Inspired by Darwin’s research on emotions (Darwin, 1872), it postulates a certain number of discrete, basic emotions that are universal and innate among humans due to their evolutionary origins. Emotions are understood as psycho-physiological "programs" that were naturally selected to help overcome recurrent evolutionary challenges (Cosmides and Tooby, 2000). A prominent version is that of Paul Ekman (Ekman, 1999), who sought to show, as Darwin envisaged, that some emotions are expressed with the same facial expressions across cultures – Ekman used Darwin’s (Darwin, 1872) list of six "core" expressions of emotions: anger, fear, surprise, disgust, happiness, and sadness. He notably conducted studies with individuals having no exposure to Western culture, indicating that they could accurately identify facial expressions for these six emotions (Ekman and Friesen, 1971). It should be noted that Ekman left it open how many basic emotions there are. Besides the six emotions listed, candidates include amusement, contempt, embarrassment, guilt, pride, and shame (Ekman, 1999). Other versions of basic emotion theory have different lists (Tomkins, 1962; Izard, 1992; Panksepp, 1998; Plutchik, 2001). For a discussion of the evidence supportive of basic emotion theory, notably the potential physiological and neurological signatures of basic emotions, see Moors (2022, 129—131).

**Psychological constructivism.** Psychological constructivism is the most influential alternative to basic emotion theory today. It rejects that there are discrete, basic emotions universally shared by humans and posits instead that emotion kinds such as anger, fear, and joy are constructed through the interplay of biological, psychological, and sociocultural factors. Early proponents include Schachter and Singer (1962), but its main representatives are James Russell and Lisa Feldman Barrett (Russell and Barrett, 1999). Psychological constructivists focus on the feeling component of emotions that they interpret as a continuum with no categorical barriers. Feelings are typically represented in a two-dimensional space with a valence axis (pleasant–unpleasant feelings) and an arousal axis (feelings of activation–deactivation). The impression that there are discrete emotions is seen as a social construct: different forms of enculturation yield different ways to conceptualize or label our bodily feelings into discrete emotional kinds. For a discussion of the evidence supportive of psychological constructivism, see Moors (2022, 261—265). Some evidence comes from so-called "arousal misattribution" studies, i.e. cases where subjects misinterpret the source of their arousal and where that seems to influence what emotions they undergo.

**Appraisal theory.** The third major psychological theory of emotion is appraisal theory, whose empirical version was pioneered by Magda Arnold (Arnold, 1960). It was developed to explain the absence of a bijective, one–to–one correspondence between kinds of emotions and emotional stimuli, i.e., the fact that the same kind of stimuli triggers different emotions and that different kinds of stimuli trigger the same kind of emotion. To explain this fact, appraisals are postulated as mediators between stimuli and emotional reactions. Appraisals are cognitive evaluations (unconscious, fast, and error-prone) of the relevance of stimuli given one’s concerns and how one should react. Appraisal theory hypothesizes that, for instance, Sam is fearful
of the mouse in the kitchen because he appraises it as an imminent threat to his safety, while Maria, on the other, is angry that there is a mouse in the kitchen because she appraises it as an intruder to be kicked out. Thus, each emotion kind can be analyzed by the associated appraisal. For instance, Lazarus (1991) proposes imminent danger for fear, demeaning offense for anger, irrevocable loss for sadness, and progress towards a goal for happiness.

In the 1980s, appraisal theorists started to analyze appraisals as regions in a multi-dimensional space (Moors et al., 2013). Appraisal dimensions typically include (a) the goal-conduciveness of the stimulus, (b) the coping potential of the individual in the situation, (c) the urgency of the needed response, (d) the cause of the eliciting event (me, others, intentional or not), and (e) the compatibility with one’s normative standards. For instance, fear is triggered by an appraisal of a stimulus as (a) highly inconducive, (b) hard to cope with, and (c) requiring an urgent response. For a discussion of the evidence supportive of appraisal theory, see Moors (2022, 190—196). Most evidence comes from self-report studies where participants are asked to recall instances of emotions and to rate these in terms of appraisal variables. Other evidence comes from manipulating appraisal dimensions and measuring associated emotions (e.g., in a video-game setting) or from neurological predictions about correlations between brain activations and appraisal dimensions.

Figure 1: The integrated framework for emotion theories. Rectangles represent the four components constituting an emotional episode, and arrows represent causation. Adapted from Scherer and Moors (2019).

An integrated framework for emotion theories. Though the three theories reviewed are usually considered rivals, some have argued for their integration (Scherer and Moors, 2019; Bonard, 2021b; Scherer, 2022). Arguably, the three theories differ mainly in their focus. Basic emotion theory focuses on the universal traits inherited from evolution, particularly their physiological and bodily expressions. Psychological constructivism focuses on the dimensions of feeling and how individuals categorize them. Appraisal theory focuses on emotional elicitation and action tendencies. We believe that a framework integrating the various elements studied by these theories is possible and desirable. What we call the integrated framework for emotion theories proposes to do so by postulating that paradigmatic emotional episodes are made of synchronized and causally interconnected changes in four components: appraisal process, action tendencies, bodily changes (motor expressions and physiological responses), and subjective feelings. For a discussion of this integrated framework, see Scherer (2022).

3 Emotion Analysis in Text

3.1 How is emotion annotated in text?

Emotion is a category. Textual emotion analysis relies on basic emotion theories to define different emotion categories to associate with textual units (a textual span, a sentence, or a document). For instance, the sentence "I love philosophy." could automatically be associated with the discrete emotion happiness. Several annotation schemes focus on subsets of categories while others encompass a broader set, reaching over 28 different categories (Demszky et al., 2020; Bostan and Klinger, 2018).

Emotion is a continuous value with affective meaning. Instead of representing emotion as a category, some annotation schemes consider emotion as a point in a multidimensional space, associating continuous values with textual units (Buechel and Hahn, 2017). These dimensions carry an affective meaning. Two dimensions dominate the literature and stem from psychological constructivism, which considers, as we have seen, that an emotion can be characterized by its degree of pleasantness and its degree of arousal. Thus, the sentence "His voice soothes me." could be automatically associated with two continuous values: a degree of pleasantness of 4 out of 5 and a degree of arousal of 1 out of 5.

Emotion is a continuous value with cognitive meaning. These dimensions can also carry a cognitive meaning. Recently, a new line of research proposes incorporating appraisal theories into emotion analysis models (Hofmann et al., 2020; Troiano et al., 2022; Zhan et al., 2023). From this perspective, emotions are caused by events
evaluated according to several cognitive dimensions. For example, the sentence "I received a surprise gift." could be automatically associated with several continuous values: the event is sudden (4 out of 5), contrary to social norms (0 out of 5), and the person has control over the event (0 out of 5).

**Emotion consists of semantic roles.** An emotion cannot be reduced to a category or continuous values with affective or cognitive meaning. To better understand an emotional event, several approaches associate spans of text with semantic roles, such as cause, target, experiencer, and cue of the emotion (Lee et al., 2010; Kim and Klinger, 2018; Bostan et al., 2020; Oberländer et al., 2020; Campagnano et al., 2022; Wegge et al., 2023; Cortal, 2024). Thus, instead of considering emotion as caused by an event, semantic role labeling of emotions considers that emotion is an event (Klinger, 2023) that must be reconstructed by answering the question: "Who (experiencer) feels what (cue) towards whom (target) and why (cause)?". In this example, each text span can be associated with a semantic role: "Louise (experiencer) was angry (cue) at Paul (target) because he did not warn her (cause)."

**Emotion is a refined feeling.** Sentiment analysis, a fundamental task in natural language processing, is sometimes considered a simplified version of emotion analysis. In its most basic form, sentiment analysis associates textual units with a category indicating a polarity (positive or negative) (Poria et al., 2020). A finer-grained task identifies aspects of a product or topic and determines the sentiment expressed about each of these aspects (Zhang et al., 2022). For example, in the sentence "The battery life of this phone is amazing, but its camera quality is disappointing.", the sentiment is positive for the aspect "battery life" and is negative for the aspect "camera quality."

### 3.2 Limitations

**No unified annotation scheme.** Divergences in the psychological definition of emotion lead to divergences in how emotion is annotated in the text. Psychological theories of emotions represent different perspectives on the emotional phenomenon. However, these perspectives are not as contradictory as they seem and may even tend towards unification (section 2). We believe this is also the case for annotation schemes in emotion analysis.

In section 5, we provide directions for constructing a unified annotation scheme inspired by recent debates in psychology (Scherer, 2022).

**Emotion verbalization is overlooked.** Emotion analysis rarely considers the process of emotion verbalization. As a result, it is difficult to obtain annotation guides that clearly define the linguistic markers to annotate in text. We want to highlight the linguistic theory of Raphael Micheli, which categorizes a broad panel of linguistic markers into three emotion expression modes (Micheli, 2014): labeled, displayed, and suggested emotion. Emotion can be expressed explicitly with an emotional label ("I am happy today"), be displayed with linguistic characteristics of an utterance such as interjections and punctuations ("Ah! That’s great!"), or be suggested with the description of a situation that, in a given sociocultural context, leads to an emotion ("She gave me a gift"). Most annotation schemes have implicitly focused on the labeled emotion, overlooking the other two expression modes. Recently, annotation schemes based on appraisal theories implicitly concern themselves with the suggested emotion (Troiano et al., 2023). Micheli’s theory thus analyzes the different types of verbal signs humans use to infer expressed emotions. In a complementary manner, theories of cognitive pragmatics are interested in the psychological mechanisms used to infer what is communicated, especially the emotions expressed by these different types of signs. In the next section, we will hypothesize that the sign categories distinguished by Micheli correspond to different sources of inferences postulated by cognitive pragmatics.

### 4 Cognitive Pragmatics and Emotional Communication

**Two analyses of communication.** Cognitive pragmatics is the branch of cognitive science concerned with how agents use and interpret signs in communication. In this and related branches, it is common to distinguish between two broad ways to analyze communication: the "dictionary analysis" (a.k.a. the "code", "semiotic", or "semantic" model) and the "detective analysis" (a.k.a. the "Gricean", "inferential", or "pragmatic" model) (Sperber and Wilson, 1995; Schlenker, 2016; Heintz and Scott-Phillips, 2023).

**Dictionary analysis.** The dictionary analysis depicts communication as a sender who intentionally
or unintentionally encodes information into a signal that the receiver decodes. Vitally, prior to the communicative exchange, the sender and the receiver must share the same code. A code here is understood as a pre-established pairing between kinds of stimuli (symbolized by "<...>") and sets of information (symbolized by "[..]"). For instance, the Morse code consists of a pairing between combinations of short and long signals and [letters] that senders and receivers must share to communicate with it. Codes can be conventional, as the Morse code is and as is the formal semantics of a language: a code made of syntactical and lexical rules that pairs [strings of words] with [sentential meanings] (Heim and Kratzer, 1998). Codes can also be non-conventional or "natural" (Wharton, 2003; Bonard, 2023a). For instance, bees are thought to use a code pairing their <dances> with the [location of nectar]. As mentioned in section 2, humans are thought to use a code pairing types of <facial expressions> with types of [emotions expressed].

The main limitation of the dictionary analysis is that codes sometimes underdetermine meaning: The pre-established pairings between <types of stimuli> and [sets of information] are sometimes insufficient to account for the information communicated. Paradigmatically, in conversational implicatures (Grice, 1975), the utterer implicitly communicates information beyond what is linguistically encoded, beyond what is determined by syntactical and lexical rules. For instance (Wilson and Sperber, 2006), if Peter asks, "Did John pay back the money he owed you?" and Mary answers, "He forgot to go to the bank.", Peter will readily understand that Mary means "no". However, the relevant code – the rules pairing <English grammar and lexicon> with [sentential meaning] – is insufficient to account for this since the code only tells you that John forgot to go to the bank.

Codes underdetermine the meaning of verbal expressions of emotions as well. To illustrate, let us go back to Micheli's typology: labeled, displayed, and suggested emotions (Micheli, 2013). As far as labeled emotions are concerned, the dictionary analysis does quite well thanks to the pairing between <emotion words> (e.g., happy, amazing, sadly) with the [emotion kinds] they refer to. However, even labeled emotions sometimes do not encode all that is communicated. For instance, "I am happy now" is explicit about the kind of emotion expressed but does not encode what the emotion is about. Nevertheless, we often correctly infer such information in the relevant context. The dictionary analysis fairs even less well with displayed emotions because these are often ambiguous. For instance, interjections such as "Wow!", "Damn!", "Fuck!", "Shit!", "Ah!", and "Oh!" though they readly display that the utterer undergoes an emotion, can express various positive and negative emotions. Furthermore, these interjections don’t encode what emotions are about. However, receivers usually correctly infer these pieces of information. The dictionary analysis regarding suggested emotions is even more limited. Depending on what the person expressing their emotion believes or desires, a phrase that only suggests emotions can communicate pretty much any kind of emotion. Imagine, for instance, that someone says, "The ship has black sails.". In a certain context, this apparently vapid sentence may poignantly convey intense emotion – because, say, it means that the son of the utterer died, as in the story of Aegaeus and Theseus. Note that, beyond verbal expression, most, if not all, types of emotional expressions also underdetermine what emotions are expressed. Facial expressions or acoustic cues (e.g., screams, laughter, sighs) also communicate different emotions given different contexts (Aviezer et al., 2008; Teigen, 2008; Vlemincx et al., 2009; Barrett et al., 2011, 2019; Bonard, 2023b). The dictionary analysis is thus also insufficient for these kinds of emotional expressions.

So, how do humans disambiguate emotional expressions in cases where codes underdetermine what is communicated? If we trust contemporary cognitive pragmatics, the answer should be found in the detective analysis of communication.

**Detective analysis.** What we call "the detective analysis" is constituted by a family of theories developed by Paul Grice (Grice, 1957, 1989) and his heirs (for reviews, see Bonard (2021a), chapter one and appendix). Note that although our presentation aims to remain balanced, no universally accepted version of this analysis exists.

As mentioned, the detective analysis was developed to account for conversational implicatures, cases where what is communicated goes beyond what is conveyed through conventional meaning, as in Peter and Mary's example above. To do so, the detective analysis conceptualizes linguistic interpretation as a type of abductive reasoning – i.e., as an inference that seeks the simplest and most
likely conclusion given the evidence available. The analysis spells out three main sources of evidence:

1. **Codes**, *i.e.*, pre-established pairings between types of stimuli and sets of information, *e.g.*, English syntactical and lexical rules; the codes for verbal and nonverbal emotional expressions. As we saw, expressions using labeled (*e.g.*, "I'm happy") and displayed emotions (*e.g.*, "Damn!") are partially understood through such codes, though they are too ambiguous to account for all that is communicated.

2. **Pragmatic expectations**, *i.e.*, how people are expected to behave in given contexts, particularly the kind of signal they receive. For instance, in conversations, people are expected to say things relevant to the question under discussion (see Grice (1975)'s maxims of conversation). For this reason, although what is literally encoded in Mary’s reply is that John forgot to go to the bank, Peter will nevertheless expect this to be relevant to the question he asked. Similarly, we expect someone’s emotional expressions to be about something relevant to their concerns (Wharton et al., 2021; Bonard, 2022). For instance, if someone says "Damn!" after receiving a surprisingly nice compliment, we expect the compliment to be particularly relevant to the person and will interpret the interjection accordingly.

3. **Common ground**, *i.e.*, the information presumed to be shared by the participants in the exchange (Stalnaker, 2002). For instance, Mary and Peter both presume that a bank is a place where one can withdraw money. Similarly, we usually presume that receiving a compliment is something that one seeks, especially if it is surprisingly nice – though this is not always part of the common ground, *e.g.*, if the complimenter is the complimentee’s arch-enemy. The common ground also allows us to understand that Aegeus can express deep despair with the sentence « The ship has black sails. ».

Based on these three sources of evidence, the detective analysis further postulates that the interpreter uses mindreading abilities (*i.e.*, theory of mind, mentalizing, or social cognition) to infer what is the most likely piece of information that is implicitly communicated – *e.g.*, Peter infers that Mary meant "no" and we infer that the person saying "Damn!" is probably pleased. Finally, the detective analysis specifies that the information so inferred is added to the common ground shared by participants in the exchange so that it may be a new source of evidence in the upcoming exchanges.

Let us note that the detective analysis predicts that the ability to correctly infer what is communicated by emotional expressions heavily depends on one’s mind-reading capacities. Corroborating this prediction, children or people on the autistic spectrum may struggle to infer implicit meaning correctly, *e.g.*, conversational implicatures (Foppolo and Mazzaggio, 2024) or in expressions using suggested emotions (Blanc and Quenette, 2017).

5 Research Directions for Emotion Analysis

5.1 Towards a Unified Annotation Scheme

Training models on data annotated with a scheme that reflects the multifaceted nature of emotions is desirable to improve the capacity of language models to understand emotions. Such a scheme would need to integrate different perspectives on the emotional phenomena to allow for better study comparisons. This would also increase the performance and generalization of models.

**Attempts at unification.** Several recent studies attempt to unify different ways of annotating emotion in text. Campagnano et al. (2022) propose a new annotation scheme that unifies various schemes on emotion semantic roles. To choose a set of shared categories, the different discrete emotions from the schemes were converted to the basic emotions of Plutchik’s theory (Plutchik, 2001). Klinger (2023) explores the divergences and commonalities between semantic role labeling of emotions and approaches based on appraisal theories. The study identifies several research directions, such as using appraisal variables to improve the task of detecting emotion causes, or analyzing experiencer-specific appraisals (Wegge et al., 2023). These studies show that combining schemes allows knowledge transfer between tasks, increasing performance and generalization.

**In search of a common framework.** What we have previously referred to as “the integrated framework for emotion theories” (section 2) aims to reconcile the main emotion theories in psychol-
In our view, it represents a strong candidate to provide a common framework for annotation schemes. As mentioned in section 2, this model considers that emotion consists of synchronized changes in different components: the appraisal process, action tendencies, bodily changes (motor expressions and physiological responses), and subjective feelings. Research in emotion analysis must draw from the recent debates in the psychology of emotions to bring existing annotation schemes into dialogue on a solid theoretical basis and, ideally, construct a unified annotation scheme.

**Emotion comprises several interacting components.** A unified annotation scheme could clarify some gray areas in emotion analysis, such as the lack of clear definitions for emotion semantic roles (e.g., experiencer, cause, and target). It could also better situate existing schemes. For example, annotating discrete emotions and affective dimensions emphasizes subjective feeling, whereas annotating cognitive dimensions emphasizes appraisals. Few schemes account for physiological responses, motor expressions, and action tendencies. More generally, few schemes consider all components. Kim and Klinger (2019) analyze the communication of emotions in fiction through descriptions of subjective sensations, postures, facial expressions, and spatial relations between characters. Casel et al. (2021) associate text spans with categories corresponding to Scherer’s emotional components. Cortal et al. (2022, 2023) structure emotional narratives according to components similar to Scherer’s. Each text span corresponds to observable behaviors, thoughts, physical feelings, or appraisals. To our knowledge, no annotation schemes attempt to capture the interaction between components. Generally, emotion analysis pays little attention to the dynamic nature of emotion and the synchronization of its various components.

**Improving the clarity of annotation guides.** We note that few studies psychologically justify the choice of different objects to detect in the text. Emotion analysis needs to develop a systematic approach to compare annotation guides with one another, thereby precisely understanding how different annotation schemes capture emotion. Thus, these schemes must draw from psychological theories (section 2) but also from linguistic theories (sections 3.2 and 4) to identify linguistic markers that verbalize emotion. With clear annotation guides, it would be easier for research teams to focus on points of convergence between schemes.

### 5.2 Better Knowledge Use and Environmental Interaction

In natural language processing, *prompting* refers to supplying a tailored input to a language model, aiming to direct its generation process towards a desired response (Brown et al., 2020). Numerous prompting methods draw inspiration from human cognition to improve the performance of language models (Zhang et al., 2023b). These methods propose generating reasoning steps (Wei et al., 2023; Kojima et al., 2023), reasoning through multiple generated responses (Wang et al., 2023b; Yoran et al., 2023), facilitating communication by rephrasing questions (Deng et al., 2023), and self-improving with its own generated feedback (Madaan et al., 2023; Yuan et al., 2024).

**Prompting methods for emotional understanding.** Most methods have been explored to improve model performance on tasks requiring formal reasoning (Zhang et al., 2023b). We believe it is possible to adapt these methods or even create new ones to improve model performance on tasks requiring social reasoning, such as emotional understanding. It would be interesting to rely on the ability of language models to act as character simulators (Shanahan et al., 2023; Lu et al., 2024), capable of adopting multiple perspectives to change style (Deshpande et al., 2023), solve tasks requiring expert knowledge (Xu et al., 2023), or simulate discussions to encourage exploration (Wang et al., 2023c; Liang et al., 2023). Zhou et al. (2023) enhance the ability of language models to make relevant inferences for solving theory of mind tasks. They propose a reasoning structure that anticipates future challenges and reasons about potential actions. More globally, a major challenge in natural language processing is finding suitable reasoning structures to effectively use the internal knowledge of models (Kojima et al., 2023; Zhou et al., 2023, 2024). The contribution of the detective analysis (section 4) could prove valuable here: prompts that explicitly ask models to seek evidence from the three sources highlighted by this analysis could lead to better performance and explainability. Finally, the integrated framework for emotion theories (section 3) can serve as inspiration for prompts that aim to exploit all the different facets of emotions rather than focusing on just one of them (e.g., subjective feeling).
Interaction with the environment. Current language models, trained solely on predicting missing words, have essentially mastered linguistic codes, i.e., lexical and syntactic rules (section 4), which Mahowald et al. (2023) call "formal linguistic competence". However, they struggle to perform well on tasks relying on what Mahowald et al. (2023) call "functional linguistic competence", i.e., the skills required to use language in real-world situations. These skills centrally involve the mechanisms postulated by the detective analysis – in particular, sharing a common ground and having sensible pragmatic expectations (section 4). To address this limitation, studies augment language models with external modules like a mathematical calculator (Schick et al., 2023), a web browser (Gur et al., 2023), or a virtual environment (Park et al., 2023). Through tool manipulation, language models intertwine reasoning with action and can thus effectively combine internal with external knowledge (Yao et al., 2023). This point is crucial to develop models that exhibit human-like social behaviors. For example, Park et al. (2023) show that observation, planning, and reflection are important components for increasing the credibility of behaviors in a virtual environment. Research on human communication can help highlight relevant abilities to augment language models (e.g., with external modules). This surely applies to emotional communication as well: models could be complemented with modules encapsulating, for instance, our knowledge of codes for emotional expressions, of how kinds of appraisals relate to kinds of emotions, and of how we expect people undergoing emotions to behave, along the lines sketched in sections 2 and 4 above.

5.3 Better Benchmarks for Emotional Understanding

Recent benchmarks evaluate language models on specific aspects of emotional understanding (Wang et al., 2023a; Paech, 2024), but they don’t consider its full richness (Scherer, 2007; Mayer et al., 2008; O’Connor et al., 2019). For example, Paech (2024) assesses emotional understanding by predicting the intensity of multiple emotions in conflict scenes. Some benchmarks evaluate models on related tasks, such as sentiment analysis (Zhang et al., 2023a) and theory of mind (Zhou et al., 2023; Ma et al., 2023; Kim et al., 2023; Gandhi et al., 2023). However, no benchmark specifically proposes to evaluate the multiple facets of emotions that affective sciences reveal (section 2). Therefore, it is difficult to know whether current models are efficient for emotional understanding.

This limitation is compounded by the fact that it is difficult to clearly determine which properties of emotional understanding are to be evaluated. We believe that evaluating language models should be grounded in research on human emotional communication, especially psycholinguistics. For example, before the age of ten, basic emotions (e.g., joy or sadness) are better remembered than complex emotions (e.g., pride or guilt) (Davidson et al., 2001; Creissen and Blanc, 2017). From six to ten years old, labeled emotions are better understood than suggested emotions (Blanc, 2010; Creissen and Blanc, 2017). Another example of relevant studies concerns the difficulty that autistic people have in understanding different types of emotional expressions (Foppolo and Mazzaggio, 2024). These studies suggest that, for humans, different types of emotions and different modes of emotional expression are more or less difficult to interpret. It would be desirable for benchmarks to evaluate language models in ways that reflect the relative difficulty of tasks for humans. Such a project would certainly benefit from research in cognitive pragmatics (section 4), knowing, for example, that people with communication disorders have difficulty understanding conversational implicatures (Foppolo and Mazzaggio, 2024), which indicates that the different sources of evidence distinguished by the detective analysis are associated with different levels of difficulty.

We believe the concept of emotion should be addressed through its relationship with text understanding, i.e., the ability of a reader to construct a mental representation of a situation in a text (Zwaan and Radvansky, 1998). Thus, we would need to go beyond current conceptualizations of emotion in natural language processing (section 3.1) to consider the diversity of linguistic markers used to verbalize emotion (section 3.2) as well as the different types of emotion (basic or complex) from psycholinguistic research (section 2). Inspired by previous studies, Etienne et al. (2022) propose an annotation scheme that considers emotion expression modes and types of emotion. Future benchmarks assessing the ability of language models to analyze emotions should consider such annotation schemes, which, as we have recommended, seek to be solidly based on relevant research in cognitive science.
6 Conclusion

Emotion analysis has several limitations that, we believe, are partially due to a lack of communication with other disciplines and, in particular, cognitive science. We propose exploiting cognitive science research on emotions and communication to address some limitations, especially what we called "the integrated framework" in emotion theories and "the detective analysis" in cognitive pragmatics. We suggest that this opens the way for constructing new annotation schemes, methods, and benchmarks for emotional understanding that consider the multiple facets of human emotion and communication.

Limitations

We propose a theoretical perspective on emotion analysis in natural language processing. We believe it would benefit the emotion analysis community to adopt an interdisciplinary approach by drawing from cognitive science theories to address certain existing limitations in the research field. In practice, this is a challenging task. Although we focus on concrete actions that could be undertaken soon (for example, clarifying annotation guidelines), we recognize that our contribution involves speculative research directions. In future research, it would be desirable to complement these speculative aspects with more concrete proposals, notably with empirically testable hypotheses and implementable algorithms.

Ethics Statement

We have not conducted any experimentation or published any data or models in this paper. The present research aims to better understand human emotional communication, not to develop tools for automatically detecting individuals’ private subjective states. While we believe our paper does not present direct ethical concerns, the research directions it raises could indirectly harm individuals and societal structures. Although we have highlighted the potential benefits of natural language processing applications (such as emotion regulation tools), it is crucial to ensure that the development and use of such tools do not have any adverse effects in the future.

Acknowledgements

Thanks to the AI-PHI group\(^1\) for making our collaboration possible and for insightful discussions. Thanks also to the ACL ARR 2024 February reviewers for useful and constructive feedback.

References


\(^1\)https://ai-phi.github.io/


Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners.


Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models.


Klaus R. Scherer. 2007. Componential emotion theory can inform models of emotional competence. Publisher: Oxford University Press.


