



**HAL**  
open science

# ”Only ChatGPT gets me”: An Empirical Analysis of GPT versus other Large Language Models for Emotion Detection in Text

Florian Lecourt, Madalina Croitoru, Konstantin Todorov

► **To cite this version:**

Florian Lecourt, Madalina Croitoru, Konstantin Todorov. ”Only ChatGPT gets me”: An Empirical Analysis of GPT versus other Large Language Models for Emotion Detection in Text. WWW '25 - Companion Proceedings of the ACM Web Conference, Apr 2025, Sydney, Australia. 10.1145/3701716.3718375 . hal-04954860

**HAL Id: hal-04954860**

**<https://hal.science/hal-04954860v1>**

Submitted on 20 Feb 2025

**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.



Distributed under a Creative Commons CC0 - Public Domain Dedication 4.0 International License

# "Only ChatGPT gets me": An Empirical Analysis of GPT versus other Large Language Models for Emotion Detection in Text

Florian Lecourt  
LIRMM UM5506 - CNRS, Université  
de Montpellier  
Montpellier, France

Madalina Croitoru  
LIRMM UM5506 - CNRS, Université  
de Montpellier  
Montpellier, France

Konstantin Todorov  
LIRMM UM5506 - CNRS, Université  
de Montpellier  
Montpellier, France

## Abstract

This work investigates the capabilities of large language models (LLMs) in detecting and understanding human emotions through text. Drawing upon emotion models from psychology, we adopt an interdisciplinary perspective that integrates computational and affective sciences insights. The main goal is to assess how accurately they can identify emotions expressed in textual interactions and compare different models on this specific task. This research contributes to broader efforts to enhance human-computer interaction, making artificial intelligence technologies more responsive and sensitive to users' emotional nuances. By employing a methodology that involves comparisons with a state-of-the-art model on the GoEmotions dataset, we aim to gauge LLMs' effectiveness as a system for emotional analysis, paving the way for potential applications in various fields that require a nuanced understanding of human language.

## CCS Concepts

• Computing methodologies → Natural language processing; Artificial intelligence.

## Keywords

Large Language Model, GPT, BERT, Emotion Detection, Emotion Model

## ACM Reference Format:

Florian Lecourt, Madalina Croitoru, and Konstantin Todorov. 2025. "Only ChatGPT gets me": An Empirical Analysis of GPT versus other Large Language Models for Emotion Detection in Text. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25)*, April 28-May 2, 2025, Sydney, NSW, Australia. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3701716.3718375>

## 1 Introduction

The advent of artificial intelligence technologies, in particular conversational agents such as ChatGPT, has profoundly transformed the way we interact with machines [25]. These agents, designed to simulate human conversations, now play a crucial role in various fields, from customer service [20] to personal assistance [40].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia*

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1331-6/2025/04

<https://doi.org/10.1145/3701716.3718375>

These novel technologies come with novel challenges for the AI community, among which we focus on one in particular: the ability to capture and correlate emotional expressions by machines and the ability of machines to express emotions and empathic behavior themselves [14].

This work aims to provide a rigorous and detailed assessment of several LLMs, including GPT and LLama, and emerging models such as Gemini, Mistral, and Phi-3, focusing on their ability to detect and respond to emotions. Given that ChatGPT became "the fastest-growing app of all time" [29], we place particular emphasis on the GPT architecture it is based on. By cross-referencing the results of different evaluation methods, we aim to identify avenues of improvement to make conversational agents more empathetic and better adapted to users' needs. Our methodological approach thus aims to answer the question "How effectively do various large language models detect and classify human emotions from text compared to a state-of-the-art emotion detection model, using macro F1 score as an objective metric?". In contexts such as mental health, customer support, and social interactions, empathy, and emotional understanding are essential [32].

The paper is structured as follows. Section 2 introduces the psychological emotion models that form the conceptual foundation for our analysis. Section 3 presents the emotion datasets used for training and evaluating AI models. Section 4 provides a quick overview of transformer-based architectures, including GPT, BERT, and other LLMs. Section 5 details our evaluation methodology, experiments, and results, covering prompt engineering techniques and cross-model comparisons. We conclude in Section 7 with comments on current findings and directions for future research.

## 2 Emotion Models

In this section of the analysis, the term **model** refers to an emotion model, as understood in the field of psychology. We begin by disambiguating terminology to avoid confusion with a possible alternative meaning in Computer Science. According to Yadollahi et al. [41], referencing the work of Fox [15], the terms emotion, mood, feeling, and affect are described in neuroscience as follows:

- **Emotion:** A discrete and consistent response to internal or external events that have a particular significance for the organism; emotion has a short-term duration.
- **Mood:** a diffuse affective state that compared to emotion is usually less intense but with longer duration.
- **Feeling:** A subjective representation of emotions, private to the individual experiencing them; similarly to emotion, it has a short-term duration.
- **Affect:** an encompassing term used to describe the topics of emotion, feelings, and moods together.

The terms **Emotion** and **Affect** are the most important here, as their uses will be found in the following works. We will now turn our attention to the various emotion models from the field of psychology. In their review of emotion models, Sreeja and Mahalakshmi [13] distinguish two categories of models:

- **Categorical (also called Discrete)** : These models feature several distinct emotions.
- **Dimensional**: These models represent emotions on continuous dimensions rather than discrete states.

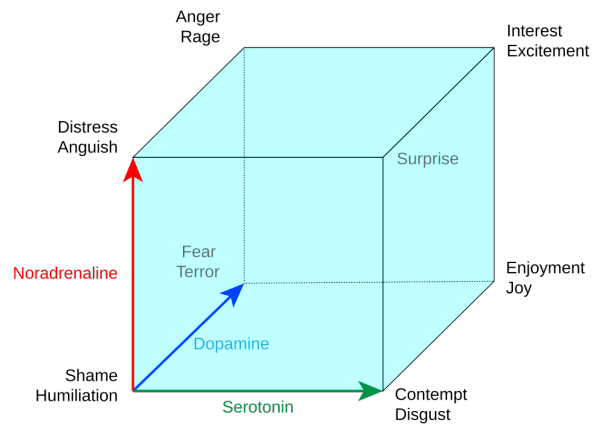


Figure 1: Graphical representation of the Lövheim model [2].

According to Yadollahi et al., "while psychologists do not agree on what model describes more accurately the set of basic emotions, the model suggested by Ekman et al., with six emotions, is the most widely used in computer science research" [41]. For Paul Ekman, this model identifies six basic emotions that are universal and recognizable by all human cultures: joy, sadness, anger, fear, surprise, and disgust [11]. Ekman developed this model from his research into facial expressions and human emotions. His first study in this domain was in 1970, where Ekman asked New Guineans to associate photographs and emotions [12]. The study's sample is of 189 adults and 130 children. Following the study's protocol, the experiment showed three photographs to a test subject, told a story concerning one of the emotions in Ekman's taxonomy, and then asked the subject to pick the photograph that fits the story. Ekman states, "The results were very clear, supporting our hypothesis that there is a pan-cultural element in facial expressions of emotion."

Before Ekman, Tomkins proposed a model comprising eight fundamental affects, identified by different facial expressions: Interest-Excitement, Pleasure-Joy, Surprise, Distress-Anguish, Fear-Terror, Shame-Humiliation, Contempt-Disgust and Anger-Rage [36]. For Tomkins, emotions "consist of one or more affects in combination with cognitive or drive states in a manner that colors, flavors, or inflects the affects" [16], corresponding to the definition we gave to the term affect. In each pair, the first term corresponds to "the most characteristic description as experienced at low [...] intensity", and the second term to the one experienced at high intensity. Tomkins used compound names for these affects to describe the expressed affect as characteristic as possible.

Building on Tomkins' work, Lövheim has developed a dimensional model represented by a cubic structure [24]. Each corner of this cube corresponds to an affect described by Tomkins. In this representation, each emotion is positioned along orthogonal axes defined by the levels of three monoamines: dopamine (DA), serotonin (5-HT), and noradrenaline (NE). For example, the Anger-Rage affect is characterized by high levels of dopamine and noradrenaline but low serotonin levels. According to Lövheim, the advantage of this dimensional model lies in its ability to correlate directly with the field of neurobiology.

Lövheim compares his approach with Plutchik's dimensional model in the introductory article to his model. Plutchik describes an eight-emotion model: Fear, Anger, Joy, Sadness, Acceptance/Trust, Disgust, Anticipation, and Surprise [27]. He justifies this choice by linking each emotion to a biological factor. In 1991, Plutchik describes an experiment in which 30 university students rate the intensity of different emotions on a scale from 1 to 11 [26]. The list includes the eight primary emotions and their synonyms. Based on the data collected, he proposes a model in which the most intense emotions are represented closer to the center and with more saturated colors than those of less intense emotions. Plutchik points out that the opposing primary emotions in this emotional wheel are complementary and that their combination produces a neutral psychic or biological state comparable to gray. The 3D version of Plutchik's model, which represents intensity on the depth axis, illustrates these concepts more explicitly.

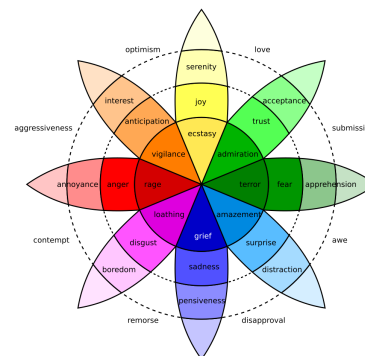


Figure 2: 2D representation of the Plutchik model [35]

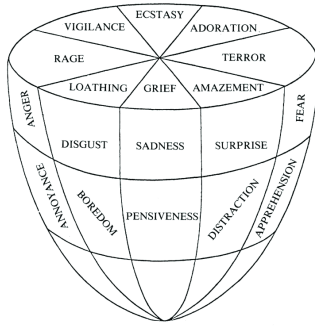


Figure 3: 3D representation of the Plutchik model [5]

	Ekman	Tomkins	Lövheim	Plutchik
Joy	✓	✓	✓	✓
Anger	✓	✓	✓	✓
Fear	✓	✓	✓	✓
Sadness	✓			✓
Acceptance/Confidence				✓
Disgust	✓	✓	✓	✓
Anticipation				✓
Surprise	✓	✓	✓	✓
Distress		✓	✓	
Shame		✓	✓	
Interest		✓	✓	
Model type	Discrete	Discrete	Dimensional	Dimensional

Table 1: Comparison of the emotions present in each model.

In Table 1, we observe that the emotions common to the different models include Joy, Anger, Fear, Disgust, and Surprise. As mentioned in the introduction of this section, there is no consensus about which model best represents the spectrum of human emotions. When these models are used in computer science, the different emotions present in each model (or affects in the case of Tomkins and Lövheim) are used to create a taxonomy to annotate datasets. While categorical models are well-suited for such a purpose, information is inevitably lost during discretization in the case of dimensional models.

### 3 Emotion Datasets

Emotion detection uses specific datasets to train and evaluate emotional classification models efficiently and accurately. The GoEmotions dataset, developed by Google, is a collection of 58,000 Reddit comments, manually annotated to cover 27 emotional categories and one neutral category [9]. This dataset stands out for its granularity, offering detailed and nuanced coverage of human emotions. Data were collected from 2005 to January 2019, excluding deleted and non-English comments. To limit bias, the data was partially filtered to reduce vulgarities while retaining those deemed essential for learning about negative emotions, limiting text length, and balancing the emotions represented. The final taxonomy of emotions was established through an iterative process to maximize the coverage of emotions expressed in the Reddit data while limiting the total number of emotions and their overlap. Initially, 56 emotional categories were considered. During iterative refinement, categories

that the annotators rarely selected showed low concordance or were difficult to detect in the text were removed to improve clarity [8]. Frequently suggested categories that were well represented in the data were added. This refinement process resulted in high annotation accuracy, with 94% of examples having at least two annotators agreeing on at least one emotional label. As a result, GoEmotions includes 12 positive, 11 negative, and four ambivalent emotions, enabling GoEmotions to serve as a reliable resource for the fine classification of emotions in texts.

The CARER dataset is a less granular dataset than GoEmotions, featuring eight emotion labels (Joy, Surprise, Anticipation, Fear, Anger, Trust, Disgust, and Sadness) [30]. Unlike GoEmotions, each text, based on tweets, is associated with a unique emotion label. The eight labels used are the same as those described by Plutchik [27]. This feature is shared by the WRIME dataset [21], composed of texts from various social networks, and GoodNewEveryone [4], which takes newspaper headlines and adds the labels Guilt, Love, Pessimism, Optimism, Pride and Shame, separating Surprise into Positive Surprise and Negative Surprise. For the latter, similarly to GoEmotions, newspaper headlines were annotated by comparing agreements between annotators.

The oldest and most cited dataset is the ISEAR dataset [31]. Based on Ekman’s work and using the emotions described in it (replacing Surprise with Shame and Guilt), ISEAR is a dataset derived from psychological research to prove the universality and cultural variations of differential emotional response patterns. The various data come from a series of questionnaires taken in 37 different countries.

Whether for datasets based on the work of Plutchik and Ekman or for GoEmotions, the shared emotions are Joy, Anger, Fear, Sadness, and Disgust. Unlike the emotions shared by the different models, the emotion of Surprise is absent here due to its non-use in the ISEAR data, and the emotion of Sadness makes its appearance, already being a common emotion in the Ekman and Plutchik models.

### 4 The Transformer models

Introduced by Vaswani et al. [38], Transformer models have revolutionized NLP thanks to their innovative architecture, overcoming the limitations of previous approaches such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) [17]. Those models are interesting notably due to their exceptional performance, which is now state-of-the-art in many fields, such as NLP [7] or audio processing [23]. Transformers are at the root of LLMs such as GPT and classifiers such as BERT. These families of models are used for emotion detection as well.

. Developed by OpenAI, GPT is an auto-regressive model [28]. This architecture generates text sequentially, predicting each subsequent word based on previously generated words. ChatGPT is a conversational agent, a chatbot, based on the GTP-3.5 model.

. In contrast, BERT is an example of an encoder model [10]. Unlike GPT, BERT is specifically designed to understand and analyze language. It excels in text classification, comprehension, and sentiment analysis tasks.

	GoEmotions	CARER	WRIME	GoodNewsEveryone	ISEAR
Admiration	✓				
Amusement	✓				
Anger	✓	✓	✓	✓	✓
Annoyance	✓			✓	
Anticipation		✓	✓		
Approval	✓				
Caring	✓				
Confusion	✓				
Curiosity	✓				
Desire	✓				
Disappointment	✓				
Disapproval	✓				
Disgust	✓	✓	✓	✓	✓
Embarrassment	✓				
Excitement	✓				
Fear	✓	✓	✓	✓	✓
Gratitude	✓				
Grief	✓				
Guilt	✓			✓	✓
Joy	✓	✓	✓	✓	✓
Love	✓			✓	
Nervousness	✓				
Neutral	✓				
Optimism	✓			✓	
Pessimism				✓	
Pride	✓			✓	
Realization	✓				
Relief	✓				
Remorse	✓				
Sadness	✓	✓	✓	✓	✓
Shame				✓	✓
Surprise	✓	✓	✓	✓	
Trust		✓	✓	✓	

Table 2: Comparison of the emotions present in each dataset.

### 4.1 Other LLMs

In this work, we investigate the emotion detection capabilities of several LLMs. In addition to GPT, we examine Gemini, Gemma, LLaMA, Phi3, Mistral, and Mixtral. Here is a short description of those models:

- **LLama [37]:** LLama models are open-source LLMs distributed by Meta. They are designed to be computationally efficient and easy to fine-tune.
- **Mistral [18]/ Mixtral [19]:** Mistral and Mixtral are two LLM introduced by Mistral AI. Mistral outperforms LLama 2 on multiple benchmarks while maintaining faster inference. Mixtral is based on a sparse mixture-of-experts (SMoE). Each token is processed by two out of eight experts per layer, giving Mixtral effective access to large parameter spaces while only using 13B active parameters per inference step. Mixtral competes with GPT-3.5 on many benchmarks.
- **Gemma [34]/ Gemini [33]:** Developed by Google, Gemma, and Gemini represent two distinct approaches in the LLM ecosystem. Gemma models are open-source solutions designed for multilingual understanding and accessibility, making them adaptable to various applications. Gemini is a multimodal LLM crafted to excel at complex benchmarks, positioning itself as a strong competitor to high-performance models like Claude 3.0 or GPT-4.

- **Phi-3 [3]:** Phi-3, a kind of model brought by Microsoft, can be described as a Small Langue Model (SLM). Despite its relatively compact size, it is designed to achieve top-tier performance and rival larger models such as Mixtral and GPT-3.5.

After establishing these models’ foundational concepts and characteristics, we can now move on to evaluating their performances in emotion detection tasks.

## 5 Evaluation and Results

### 5.1 Emotion Detection

Natural Language Processing (NLP) is an essential branch of artificial intelligence devoted to understanding and manipulating human language by machines. NLP problems can be divided into two main categories: symbolic and statistical [39]. Statistical approaches are the basis behind Transformer models and LLMs, so we focus on these methods here.

While traditional opinion mining, or sentiment analysis, classifies opinions as positive, negative, or neutral, emotion detection (ED) offers a more nuanced understanding of affective states [6]. By moving beyond a binary or ternary scale, ED captures subtle emotional cues, paving the way for more empathetic and contextually aware AI applications.

### 5.2 Chat-GPT and Emotion Detection

After exploring Transformers models, BERT, LLMs, and emotion datasets such as GoEmotions, it is pertinent to look at the comparative evaluation of these models in the specific domain of emotion detection. The article *ChatGPT: Jack of all trades, master of none* evaluates ChatGPT’s performance on various NLP tasks, including emotion detection [22]. This evaluation compares ChatGPT with models considered to be state-of-the-art (SOTA) for the same tasks.

In the field of emotion detection, ChatGPT is evaluated as a classifier. Its performance is measured using the GoEmotions dataset. Given the variability in the numbers of each emotional class in this dataset, the F1 macro score is used as the evaluation metric. The F1 macro score is calculated as the arithmetic mean of the individual F1 scores for each class, where each F1 score is itself the harmonic mean of precision and recall for that class. This method enables a balanced evaluation by not favoring any particular class, regardless of their prevalence in the dataset. This property is essential in contexts where classes are unequally represented, as it prevents the bias towards majority classes that could distort the overall assessment of model performance. By balancing the influence of each class, the F1 macro encourages the development of models that effectively recognize all emotions, including less frequent ones, thus contributing to a richer understanding of the emotional nuances captured in the text.

### 5.3 Reproduction of Results

In the following section, we specifically seek to reproduce the results observed in Kocon’s study [22] to verify the consistency of ChatGPT’s performance in emotion detection, as described in this publication.

Firstly, the BERT model, referred to as SOTA, used by the article’s authors, is tested to confirm its F1 macro score [1]. The second step is to use the OpenAI API to interact with GPT-3.5-Turbo, which is identical to the one on which ChatGPT is based. A specific prompt is sent through the API to evaluate ChatGPT, which then generates the model response. The structure of this prompt is inspired by the article, as illustrated in Figure 4. The response received from ChatGPT is then analyzed to calculate its F1 macro score. The BERT model and GPT-3.5-Turbo will be tested using the **test** set from the dataset GoEmotions.

Evaluated metrics include:

- **ChatGPT macro F1 score (%)**: Calculated as the average of the F1 scores for each class, this measures ChatGPT’s overall performance across all classes regardless of their frequency of appearance.
- **SOTA macro F1 score (%)**: Measures the performance of the SOTA model for the same task. Calculated in the same way as ChatGPT’s F1 macro.
- **Difference (pp)**: The difference in percentage points between the F1 macro scores of ChatGPT and the SOTA model.
- **Difficulty (%)**: Defined as

$$\text{Difficulty} = 100\% - F1_{\text{macro, SOTA}}$$

This metric reflects the task’s intrinsic difficulty based on the SOTA model’s performance.

- **Loss (%)**: Calculated as

$$\text{Loss} = 100\% \times \frac{F1_{\text{macro, SOTA}} - F1_{\text{macro, ChatGPT}}}{F1_{\text{macro, SOTA}}}$$

This metric shows the performance loss of ChatGPT compared with the SOTA model.

Chat 58. Task: GoEmo. Case 7.
<b>Prompt</b>
From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in the form of a Python list containing exactly 1 selected most matching emotion. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. Text: <i>You're welcome.</i>

Figure 4: Example prompt [22]

Analysis of ChatGPT’s performance in comparison with the SOTA model on the emotion detection task, as illustrated in Table 3, reveals a significant deviation from the performance of the SOTA model. This discrepancy is noticeable in all the contexts tested, with a performance loss of more than 50% in all contexts. This observation suggests that, despite ChatGPT’s advanced text generation capabilities, its performance in the specific emotion detection task remains substantially inferior to that of a model dedicated to this task, confirming the article’s conclusions. The various tests were carried out with varying batch sizes due to the constraints imposed by the OpenAI API. In the following section, the batch size used

for testing will be the one from **Test2**, as the results obtained for this test are the closest to the one in Kocon’s paper.

### 5.4 Evaluation Setting

The methodology of this study consists of several steps aimed at evaluating and improving the performance of ChatGPT for the emotion detection task. First, we thoroughly review prompt engineering techniques, building on approaches identified in the state of the art to optimize the instructions given to ChatGPT. The aim is to maximize its F1 macro score, a metric chosen to evaluate the model’s accuracy in a balanced way across all emotional classes.

Once the best prompt has been determined, we will compare ChatGPT’s performance with other language models using the same optimized prompt. Once again, ChatGPT is represented by the GPT-3.5-turbo model, on which it is based. This comparison will enable us to situate ChatGPT in relation to other models in the specific context of emotion detection. Then, to check the results’ robustness, we will employ complementary methods, such as integrating dictionaries to correct responses that do not appear in the list of 28 emotions. Figure 5 shows a flowchart of the evaluation.

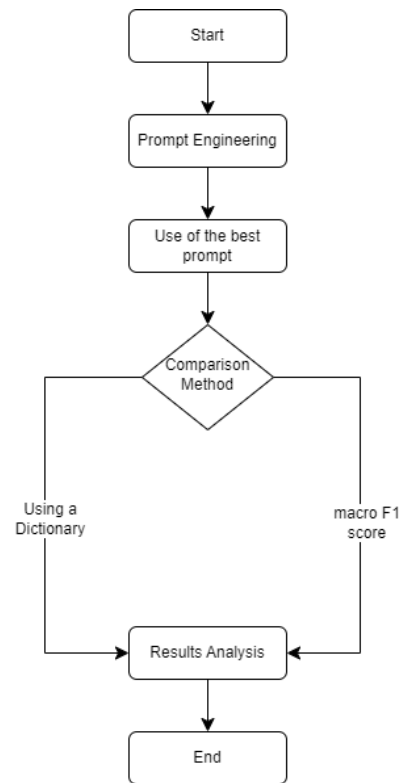


Figure 5: Evaluation flowchart

### 5.5 Prompt Engineering

To optimize GPT’s performance in the emotion detection task, we explored several variants of prompts. Each variant aims to refine the instructions given to the model to improve the accuracy and consistency of responses. The four prompts used in this study are

	ChatGPT macro F1 score (%)	SOTA macro F1 score (%)	Difference (pp)	Difficulty (%)	Loss (%)
Reference values	25.55	52.75	27.20	47.25	51.56
Test1, Batch size 500	22.43	48.86	26.43	51.14	54.09
Test2, Batch size 1000	22.82	52.19	29.37	47.81	56.28
Test3, Batch size 2500	22.83	49.30	26.47	50.70	53.69
Test4, Entire dataset	23.02	49.68	26.66	50.32	53.67

Table 3: Comparison of Chat-GPT and SOTA model performance depending on the batch size.

detailed below, each with specific adjustments to maximize the F1 macro score. The basic prompt (Figure 6) asks GPT to select a single emotion from a given list elicited by the text provided. This prompt serves as a starting point for evaluating the initial performance of the GPT model.

```
prompt = (f "From the given list of all emotions, choose 1 emotion
↳ that the input text arouses in most people reading it. Write
↳ your answer in a Python list. List of all emotions: {'',
↳ '.join(emotion_labels)}. Input text: {text}")
```

Figure 6: Original prompt

The new prompt, seen in Figure 7, adds a variable for the number of emotions to be identified, corresponding to the number of emotions annotated for the given text in the GoEmotions dataset. This approach better aligns GPT’s responses with the dataset’s annotations.

```
prompt = (f "From the given list of all emotions, choose
↳ {number_of_emotions} emotions that the input text arouses in
↳ most people reading it. Write your answer in a Python list.
↳ List of all emotions: {'', '.join(emotion_labels)}. Input text:
↳ {text}")
```

Figure 7: First Variant

For the next prompt, emphasis is placed on the exact number of emotions to be returned using the phrase "Please list exactly number\_of\_emotions." This formulation is intended to reduce ambiguity and encourage GPT to adhere strictly to the requested number of emotions. This prompt is illustrated in Figure 8.

```
prompt = (f "Please list exactly {number_of_emotions} emotions that
↳ the following text arouses in most people, separated by commas.
↳ Write your answer in a Python list. Use the emotions from the
↳ following list only: {'', '.join(emotion_labels)}. Text:
↳ '{text}'")
```

Figure 8: Second Variant

The last prompt (Figure 9) repeats the methodology of the previous prompt while adding quotation marks around the number

of emotions requested and providing an explicit example of the expected response format. This example is intended to clarify expectations further and guide GPT towards a correctly formatted response.

```
prompt = (f "Please list exactly \" {number_of_emotions}\" emotions
↳ that the following text arouses in most people, separated by
↳ commas. Respond with your answer in a Python list, selecting
↳ only from the provided list of emotions. For instance, if the
↳ text evokes fear and excitement in most people and you are
↳ asked for 2 emotions, you should write: ['fear', 'excitement'].
↳ If the request is for a single emotion, select the most
↳ dominant, like this: ['fear']. Refer exclusively to the
↳ following list of valid emotions: {'', '.join(emotion_labels)}.
↳ Based on this list only, what are the \"{number_of_emotions}\"
↳ emotions that this text arouses: '{text}'")
```

Figure 9: Third Variant

	Model macro F1 score (%)	Difference (pp)	Loss (%)
Reference values	22.82	29.37	56.28
Variant 1	27.28	24.91	47.73
Variant 2	26.14	26.05	49.91
Variant 3	28.97	23.22	44.49

Table 4: Comparison of Chat-GPT performance depending on the prompt used.

Table 4 shows that the last prompt achieves the highest F1 macro score. In the remainder of this study, we will use this prompt to explore Chat-GPT’s performance in greater depth and compare it with other language models.

### 5.6 Comparisons with other LLMs

As mentioned in the previous subsection, we will now compare the F1 macro scores of Chat-GPT with those of other large language models (LLMs). The aim is to determine whether one model outperforms GPT-3.5-Turbo in the emotion detection task. The prompt ??, which gave the best results for Chat-GPT, will be used for these comparisons.

Gemini-1.5 results were obtained using a Google Colab provided by Google. The performance of Llama-3-70b and Mixtral-8x7b was measured via the Huggingchat API, as these models are too large to be run locally. The other results were obtained by running the models locally using the Ollama application and Python library.

Model name	Model macro F1 score (%)	Difference (pp)	Loss (%)
GPT-3.5-Turbo	28.97	23.22	44.49
GPT-4o	30.95	21.24	40.70
Llama-2-7b	20.24	31.95	61.22
Llama-3-8b	20.60	31.59	60.53
Llama-3-70b	27.20	24.99	47.88
Phi-3-4k	25.23	26.96	51.66
Gemma-1.1-7b	22.89	29.30	56.14
Gemma-2-9b	24.21	27.98	53.61
Gemini-1.5	26.74	25.45	48.76
Mistral-7b	25.14	27.05	51.83
Mixtral-8x7b	23.82	28.37	54.36

**Table 5: Comparing the performance of different language models.**

Analysis of the results presented in Table 5 reveals significant differences between language model families. Models in the GPT family, including GPT-3.5-Turbo and GPT-4o, stand out for their overall superior performance in emotion detection. In particular, GPT-4o shows a slight improvement over GPT-3.5-Turbo, underlining the continued progress in this series.

The Llama family models, particularly Llama-3-70b, also show promising skills, albeit slightly inferior to those of the GPT models. Lighter versions, such as Llama-2-7b and Llama-3-8b, do not achieve the same level of performance, indicating a correlation between model size and emotion detection capabilities for this model family.

Google-developed models, such as Gemini-1.5, Gemma-1.1-7b, and Gemma-2-9b, show respectable results, although they do not surpass GPT models. However, this model family continues to offer a solid alternative with consistent performance.

The Mistral and Mixtral models show less competitive results compared to the GPT and Llama-3-70b models, although they have superior skills compared to the other Llama models.

Finally, the Phi-3 model, developed by Microsoft, shows competitive performance, positioning itself between the Llama and Google models regarding the macro F1 score. Phi-3 is a Small Language Model (SLM), a category of models developed by Microsoft to offer capabilities similar to those of large language models but with reduced size and resource requirements. SLMs thus offer an efficient alternative to LLMs for specific tasks.

In summary, GPT models dominate in terms of performance, followed by Llama and Google models. Though inferior performers, Mistral, Mixtral, and Phi-3 may offer viable alternatives.

Although the macro F1 score or Accuracy are widely used and enables a standardized performance comparison between different models, they have certain limitations when it comes to capturing the subtlety of the errors made by these models. In particular, they treat each error equally without considering the semantic proximity between predicted and true emotions. This binary approach to errors is problematic in emotion detection, where certain emotions are intrinsically closer to each other, especially in fine-granulated datasets such as GoEmotions.

### 5.7 Using a Dictionary

ChatGPT and the other LLMs sometimes respond outside the requested emotions list. In the previous results, these responses were

treated as 'neutral'. We will test a new approach of reclassifying these incorrect responses into the correct tags to see if this increases the scores of the different models. To do this, we will use the spaCy library, which specializes in natural language processing problems.

Using the different SpaCy models (SM, MD, and LG) and the `similarity()` function included, we created a function that takes as input an incorrect response and the tag list and returns as output the tag predicted with the highest semantic similarity to the incorrect response. To better observe the differences between using this approach and the approach without the use of dictionaries, the Table 6 will show for each model, in each case, the macro F1 scores, precision, recall, and accuracy obtained, with a precision of five decimal places.

Model name	Result type	Dictionary size	Model macro F1 score (%)	Precision (%)	Recall (%)	Accuracy (%)
GPT-3.5-Turbo	Original With Dictionary	N/A	28.97477	31.96756	36.73094	23.71250
		SM	28.47332	31.96914	36.70862	22.77500
		MD	28.66473	31.15335	37.59664	22.93750
		LG	28.55063	30.71852	37.66528	22.97500
GPT-4o	Original With Dictionary	N/A	30.94547	32.48368	38.20576	24.10000
		SM	30.64027	31.99912	38.18851	23.28750
		MD	30.64368	32.50769	38.50577	23.31250
		LG	30.65712	32.25237	38.63715	23.37500
Gemini-1.5	Original With Dictionary	N/A	26.74038	33.01710	31.71704	20.48750
		SM	26.65361	32.36802	31.74466	20.25000
		MD	26.80581	32.84068	32.05650	20.21250
		LG	26.84648	32.70287	32.13538	20.23750
Gemma-1.1-7b	Original With Dictionary	N/A	22.89399	33.10732	29.47269	22.66250
		SM	21.70670	32.57787	29.29456	18.15000
		MD	22.91514	26.63251	31.50328	18.98750
		LG	22.47977	27.28477	31.63337	18.98750
Gemma-2-9b	Original With Dictionary	N/A	24.20599	31.16082	31.66164	17.93750
		SM	23.95058	30.88174	31.65361	17.27500
		MD	24.21847	30.79365	32.42775	17.32500
		LG	24.37492	31.02687	32.64281	17.36250
Llama-2-7b	Original With Dictionary	N/A	20.24248	36.05048	21.86551	10.48750
		SM	19.60433	35.70024	21.96700	9.42500
		MD	20.72441	28.03009	24.03709	9.33750
		LG	20.60805	27.94807	24.37379	9.33750
Llama-3-8b	Original With Dictionary	N/A	20.59527	30.72653	27.36376	16.15000
		SM	20.12037	29.84229	27.33743	14.61250
		MD	20.50476	27.82015	28.13073	14.85000
		LG	20.31532	26.90405	28.27545	14.90000
Llama-3-70b	Original With Dictionary	N/A	27.20091	34.85671	33.62490	22.61250
		SM	26.96111	34.65644	33.65964	22.07500
		MD	27.22124	34.01447	34.32349	22.16250
		LG	27.25740	33.97677	34.41948	22.18750
Phi-3-4k	Original With Dictionary	N/A	25.23149	26.76328	37.07785	12.35000
		SM	24.79894	26.96175	37.13824	11.93750
		MD	25.25119	27.53842	37.91341	12.05000
		LG	26.01663	26.75966	38.70548	12.13750
Mistral-7b	Original With Dictionary	N/A	25.14239	28.64566	33.75372	18.97500
		SM	24.14168	28.36949	33.52505	16.41250
		MD	24.69657	26.09023	35.59150	16.68750
		LG	24.82202	25.76153	35.85328	16.75000
Mixtral-8x7b	Original With Dictionary	N/A	23.81649	28.67170	29.64733	19.63750
		SM	22.65727	27.31309	29.37475	15.43750
		MD	16.22614	12.88130	35.40475	14.76250
		LG	15.81165	12.31099	35.81679	14.77500

**Table 6: Comparison of the performance of different language models based on the use of dictionaries of different sizes**

Analysis of the results presented in Table 6 shows that using dictionaries to reclassify incorrect responses has a variable impact on the performance of the different language models. Integrating dictionaries generally reduces macro F1 score and precision but improves recall. This drop in macro F1 score can be explained by invalid responses no longer being classified under the Neutral tag, further reducing the number of responses in this category. Large language models (LLMs) often have difficulty giving Neutral as an



answer, so some of the wrong answers are counted as true Neutral positives.

For example, while improving recall, the GPT models see a notable decrease in precision when a dictionary is used. Similarly, the Mistral and Phi-3 models show similar trends, where the improvement in recall does not compensate for the loss in precision and macro F1 score. These observations confirm that the dictionary-based approach to correcting incorrect responses is not optimal for improving the overall performance of language models in emotion detection.

This method will not be used in the future, as automatic synonym search is an open problem that does not yield satisfactory results. The disparity in scores between different model sizes highlights the limitations of this approach, with performance varying significantly between small, medium, and large models.

## 6 Limitations

While this study provides insights into LLMs' emotion detection capabilities, its primary reliance on the GoEmotions dataset displays limitations concerning generalizability. To further validate our findings, future research should explore their validity across datasets with diverse structures and assess model robustness against varying annotation schemes and cultural contexts. Incorporating benchmarks from multiple datasets could further validate our conclusions.

## 7 Conclusion and Future Work

In this work, we investigated the capabilities of LLMs in detecting and understanding human emotions through text, aiming to improve human-computer interaction by making AI technologies more responsive to emotional nuances. While we focused on statistical approaches using the GoEmotions dataset, we acknowledge that evaluating multiple datasets would strengthen the generality of our findings. Although ChatGPT and other LLMs demonstrate advanced text generation capabilities, their performances in emotion detection remain inferior to specialized models. However, applying prompt engineering techniques brought significant improvements, highlighting the importance of subtle guidance in eliciting more accurate responses. While LLMs may not surpass specialized classifiers like BERT in emotion detection tasks, the insights from this comparative study provide a valuable foundation for refining their performance.

Looking ahead, future efforts include introducing a new evaluation metric that accounts for semantic proximity between predicted and true emotions, rewarding near-correct predictions, and penalizing distant ones. Constructing a dedicated dialogue corpus would also allow more precise testing of a model's adaptability to linguistic and emotional nuances. Furthermore, future work will incorporate rigorous statistical validation to ensure that observed performance differences between models are statistically significant and not due to random chance.

In conclusion, our research highlights the strengths and weaknesses of LLMs in emotion detection. Continuing this work could contribute to the evolution of artificial intelligence technologies, leading to a better understanding and a more empathetic response to human emotions.

## Acknowledgments

The authors gratefully acknowledge the financial support provided by the European fund for regional development FEDER through the IA-EMOTIONS project.

## References

- [1] [n. d.]. Monologg/Bert-Base-Cased-Goemotions-Original · Hugging Face. <https://huggingface.co/monologg/bert-base-cased-goemotions-original>
- [2] 2024. Lövhheim Cube of Emotions. *Wikipedia* (July 2024).
- [3] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* (2024).
- [4] Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. [n. d.]. GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (Marseille, France, 2020-05), Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, 1554–1566. <https://aclanthology.org/2020.lrec-1.194>
- [5] Patricia Bota, Chen Wang, Ana Fred, and Hugo Plácido da Silva. 2019. A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals. *IEEE Access PP* (09 2019), 1–1. <https://doi.org/10.1109/ACCESS.2019.2944001>
- [6] Mondher Bouazizi and Tomoaki Ohtsuki. [n. d.]. Sentiment Analysis: From Binary to Multi-Class Classification: A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter. In *2016 IEEE International Conference on Communications (ICC)* (Kuala Lumpur, Malaysia, 2016-05). IEEE, 1–6. <https://doi.org/10.1109/ICC.2016.7511392>
- [7] Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. [n. d.]. *Transformers: "The End of History" for NLP?* <https://doi.org/10.48550/arXiv.2105.00813> arXiv:2105.00813
- [8] Dorotyya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. [n. d.]. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, 2020-07), Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics. <https://research.google/blog/goemotions-a-dataset-for-fine-grained-emotion-classification/>
- [9] Dorotyya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547* (2020).
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. [n. d.]. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. <https://arxiv.org/abs/1810.04805> (n. d.). arXiv:1810.04805 <http://arxiv.org/abs/1810.04805>
- [11] Paul Ekman. [n. d.]. An Argument for Basic Emotions. 6, 3-4 (n. d.), 169–200. <https://doi.org/10.1080/02699939208411068>
- [12] Paul Ekman and Dacher Keltner. 1970. Universal facial expressions of emotion. *California mental health research digest* 8, 4 (1970), 151–158.
- [13] given-i=SREEJA family=P S, given=SREEJA and Mahalakshmi G S. [n. d.]. Emotion Models: A Review. 10 (n. d.), 651–657.
- [14] Alois Ferscha. 2016. A research agenda for human computer confluence. *Human Computer Confluence Transforming Human Experience Through Symbiotic Technologies* (2016), 7–17.
- [15] Elaine Fox. [n. d.]. *Emotion Science: Cognitive and Neuroscientific Approaches to Understanding Human Emotions*. <https://doi.org/10.1007/978-1-137-07946-6>
- [16] Adam J Frank and Elizabeth A Wilson. 2020. *A Silvan Tomkins handbook: Foundations for affect theory*. U of Minnesota Press.
- [17] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. [n. d.]. Overview of the Transformer-based Models for NLP Tasks. 179–183. <https://doi.org/10.15439/2020F20>
- [18] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
- [19] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024).
- [20] Alekya Jonnal. [n. d.]. How Large Language Models (LLM) Help Enterprises Enhance Customer Experiences. 13, 11 (n. d.).
- [21] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. [n. d.]. WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations. In *Proceedings of the 2021 Conference*

- of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologies* (Online, 2021-06). Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 2095–2104. <https://doi.org/10.18653/v1/2021.naacl-main.169>
- [22] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniec, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. [n. d.]. ChatGPT: Jack of All Trades, Master of None. 99 ([n. d.]), 101861. <https://doi.org/10.1016/j.inffus.2023.101861>
- [23] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. [n. d.]. Efficient Training of Audio Transformers with Patchout. In *Interspeech 2022* (2022-09). 2753–2757. <https://doi.org/10.21437/Interspeech.2022-227>
- [24] Hugo Lövhelm. [n. d.]. A New Three-Dimensional Model for Emotions and Monoamine Neurotransmitters. 78 ([n. d.]), 341–8. <https://doi.org/10.1016/j.mehy.2011.11.016>
- [25] Bojan Obrenovic, Xiao Gu, Guoyu Wang, Danijela Godinić, and Ilimdorjon Jakhongirov. [n. d.]. Generative AI and Human-Robot Interaction: Implications and Future Agenda for Business, Society and Ethics. ([n. d.]). <https://doi.org/10.1007/s00146-024-01889-0>
- [26] Robert Plutchik. [n. d.]. *The Emotions*. University Press of America.
- [27] R Plutchik. 1982. A psycho evolutionary theory of emotions. *Social Science Information* (1982).
- [28] Alec Radford. 2018. Improving language understanding by generative pre-training. (2018).
- [29] Jürgen Rudolph, Shannon Tan, and Samson Tan. 2023. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching* 6, 1 (2023), 364–389.
- [30] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. [n. d.]. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium, 0010/2018-11). Association for Computational Linguistics, 3687–3697. <https://doi.org/10.18653/v1/D18-1404>
- [31] Klaus R. Scherer and Harald G. Wallbott. [n. d.]. Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning. 66, 2 ([n. d.]), 310–328. <https://doi.org/10.1037/0022-3514.66.2.310>
- [32] Ruosi Shao. 2023. An Empathetic AI for Mental Health Intervention: Conceptualizing and Examining Artificial Empathy. In *Proceedings of the 2nd Empathy-Centric Design Workshop*. 1–6.
- [33] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).
- [34] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295* (2024).
- [35] From Wikipedia the free encyclopedia. [n. d.]. Robert Plutchik. [https://en.wikipedia.org/w/index.php?title=Robert\\_Plutchik&oldid=1240659436](https://en.wikipedia.org/w/index.php?title=Robert_Plutchik&oldid=1240659436)
- [36] Silvan Tomkins. [n. d.]. *Affect Imagery Consciousness: Volume I: The Positive Affects*. Springer Publishing Company.
- [37] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [38] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
- [39] Stefan Wermter, Ellen Riloff, and Gabriele Scheler. [n. d.]. *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*. Springer Science & Business Media.
- [40] Zhenyu Xu, Hailin Xu, Zhouyang Lu, Yingying Zhao, Rui Zhu, Yujiang Wang, Mingzhi Dong, Yuhu Chang, Qin Lv, Robert P Dick, et al. [n. d.]. Can Large Language Models Be Good Companions? An LLM-based Eyewear System with Conversational Common Ground. ([n. d.]). [arXiv:2311.18251](https://arxiv.org/abs/2311.18251)
- [41] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R. Zaiane. [n. d.]. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. 50, 2 ([n. d.]), 1–33. <https://doi.org/10.1145/3057270>