Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions

Flor Miriam Plaza-del-Arco\textsuperscript{1}, Alba Curry\textsuperscript{2}, Amanda Cercas Curry\textsuperscript{1}, Dirk Hovy\textsuperscript{1}

\textsuperscript{1}MilaNLP, Bocconi University, Department of Computing Sciences, Milan, Italy
\textsuperscript{2}School of Philosophy, Religion and History of Science, University of Leeds
{flor.plaza, amanda.cercas, dirk.hovy}@unibocconi.it
a.a.cercascurry@leeds.ac.uk

Abstract

Emotions are a central aspect of communication. Consequently, emotion analysis (EA) is a rapidly growing field in natural language processing (NLP). However, there is no consensus on scope, direction, or methods. In this paper, we conduct a thorough review of 154 relevant NLP publications from the last decade. Based on this review, we address four different questions: (1) How are EA tasks defined in NLP? (2) What are the most prominent emotion frameworks and which emotions are modeled? (3) Is the subjectivity of emotions considered in terms of demographics and cultural factors? and (4) What are the primary NLP applications for EA? We take stock of trends in EA and tasks, emotion frameworks used, existing datasets, methods, and applications. We then discuss four lacunae: (1) the absence of demographic and cultural aspects does not account for the variation in how emotions are perceived, but instead assumes they are universally experienced in the same manner; (2) the poor fit of emotion categories from the two main emotion theories to the task; (3) the lack of standardized EA terminology hinders gap identification, comparison, and future goals; and (4) the absence of interdisciplinary research isolates EA from insights in other fields. Our work will enable more focused research into EA and a more holistic approach to modeling emotions in NLP.

Keywords: emotion analysis, survey, trends, gaps.

1. Introduction

Emotions perfume our every experience and interaction, playing a key role in human cognition and relationships. In recent decades, research on emotion has become popular in numerous fields, including psychology, humanities, and social sciences. In natural language processing (NLP), interest in emotion mostly translates as emotion analysis (EA), whose growth has been most notable since 2018 (Strapparava and Mihalcea, 2008; Mohammad et al., 2018; Klinger et al., 2018; Nissim et al., 2020; Mohammad, 2021; Barnes et al., 2022).

We survey\textsuperscript{1} over 150 ACL papers (2014-2022)\textsuperscript{2} on EA to address four questions: (1) How are EA tasks defined in NLP? (2) What are the most prominent emotion frameworks, and which emotions are modeled? (3) Is the subjectivity of emotions considered in terms of demographics and cultural factors? (4) What are the primary NLP applications for EA?

We find a lack of consensus in terms of the scope, direction, and methods in the current studies, which raises three important concerns:

1. The absence of demographic and cultural aspects does not account for the variation in how emotions are presented and perceived but assumes they are universally experienced in the same manner. There is a general lack of engagement with recent work on how emotions are racialized (Judd, 2019), gendered (Campbell, 1994), etc. In other words, who produces the data matters (see Figure 1) and who annotates it.

2. The poor fit of emotion categories to the task poses a significant challenge. The commonly used predefined emotion categories may not adequately capture the nuances required for the

![Figure 1: Variation in emotion annotation based on demographics. Annotators with distinct demographic profiles, a 60-year-old male and a 20-year-old male, exhibit varying interpretations of the emotions evoked.](image-url)
downstream task.

3. The lack of a common systematic nomenclature in EA obscures gaps, limits comparison, and, therefore, future objectives. For example, our survey shows that ‘prediction’ is synonymous with ‘classification.’ In other areas of NLP, these two terms are used for distinct tasks.

4. The dearth of interdisciplinary research means progress in other fields does not inform EA. A main problematic area is at the level of task design: NLP still predominantly relies on the work of Paul Ekman (Ekman and Friesen, 1981). Ekman’s work addressed the question of basic emotions, not to provide a comprehensive account of emotions nor to define the most ‘helpful’ emotions. Thus, the categories this theory provides are too broad and oversimplified to be useful across tasks and languages (De Bruyne, 2023).

Emotions are crucial in real-life applications. This paper also explores the diverse range of emotion applications, focusing on dialogue emotion recognition as the most prominent and widely used. We analyze widely used datasets tailored for this purpose, highlighting their key characteristics.

Unlike previous surveys in the field, ours focuses on unexplored questions, revealing critical gaps in the EA field that require community attention. We aim to connect these gaps, fostering focused research and nuanced emotion modeling in NLP. We conclude with recommendations and suggestions for future work.

2. Related Work

EA has attracted considerable research interest. Various surveys provide a comprehensive overview of the state of the field (Canales and Martínez-Barco, 2014; Yadollahi et al., 2017; García-García et al., 2017; Mäntylä et al., 2018; Kim and Klinger, 2018a; Saxena et al., 2020; Acheampong et al., 2020). Recent surveys (Murthy and Kumar, 2021; Kusal et al., 2022b; Singh Tomar et al., 2023) have primarily focused on the identification of datasets, models, detection techniques, affective computing modalities (visual, vocal, textual, etc.), and applications in the literature. While these surveys acknowledge certain limitations in EA, such as the scarcity of labeled datasets, data imbalance, the dominance of English, and the performance of NLP methods, most of them do not address the significant gaps we discuss. Specifically, most disregard the significance of demographic factors in their analysis (with a brief mention of the work by Sailunaz et al. (2018)), aggregation methods for collecting annotator perspectives, or the importance of interdisciplinary.

Mohammad (2022) conducted a comprehensive analysis of AI Ethics and Emotion Recognition literature, synthesising 50 ethical considerations relevant to automatic emotion recognition. His work provides valuable ethics pointers that consider some topics we covered here, such as the crucial consideration of demographic factors and the modeling of annotators’ perspectives during dataset creation. The gaps we identified demonstrate the inherent connection between ethics and advancements in NLP for EA.

3. Survey of Emotion Analysis in NLP

In this section, we present our survey of papers from the ACL Anthology. To ensure a comprehensive selection of studies on EA, we focused on identifying papers whose titles or abstracts include keywords associated with EA NLP tasks in the field. They are ‘emotion analysis’, ‘emotion prediction’, ‘emotion classification’, ‘emotion detection’, ‘emotion recognition’, ‘emotion polarity classification’, and ‘emotion cause detection’. As EA increased in frequency after 2018, we doubled the time horizon to consider papers published in the last ten years. We removed duplicates and papers from shared task participants, resulting in 438 studies. We applied additional criteria to ensure the inclusion of influential and relevant works in our survey. Specifically, we opted for papers with a citation count from Semantic Scholar surpassing the median average (excluding self-citations). However, we adjusted this threshold to at least 1 citation for papers published in the last and current year. Following this selection process, our final analysis set contains 154 papers. Among these, 119 are published in main conferences, 31 in workshops, 3 in Findings, and 1 in other venues. For each paper, we collected the following information: Emotions used, emotion model type, emotion model, language, resource (dataset, lexicon), data source, presence of multimodality (yes, no), nomenclature used, and applications. If a paper focuses on dataset creation, we also determined whether the annotators’ demographics were considered.

Figure 2 provides an overview of the 154 papers by year and keyword. There is a clear trend in the number of papers from 2014 to 2022, with 2018 marking a significant increase in the appearance of EA studies. This jump indicates a growing interest in the field during the last five years.

At the time of this study (September 2023), the ACL Anthology contained 88,294 papers, including those from non-ACL events like LREC.

While we acknowledge the existence of other relevant terms in EA, most papers pertinent to our core questions use predominantly these terms.

We provide the survey data on GitHub: https://github.com/MilaNLProc/emotion_analysis_survey.
Figure 2: Distribution of papers considered in our survey by year and keyword.

3.1. Nomenclature

Figure 2 also provides an overview of the commonly used terms to describe EA tasks. The term ‘emotion recognition’ stands out as the most frequently used (42.48%), followed by ‘emotion analysis’ (28.10%), emotion classification (13.07%), and ‘emotion detection’ (12.41%). ‘Emotion prediction’ and ‘emotion cause detection’ are the least used terms.

However, during our analysis, we note that these terms are not used in isolation. Figure 3 highlights the most prevalent pairs of terms used across different papers to refer to these tasks. The most frequent pair, ‘classification-recognition,’ is used interchangeably across the papers, followed by ‘classification-analysis’ and ‘classification-detection.’ This variation indicates a lack of standardized terminology for EA tasks.

Given the lack of uniformity, providing definitions for each term is challenging. We have chosen EA as the broad category that demarcates any emotion-related tasks in NLP. It should be distinguished from sentiment analysis or opinion mining. There is consensus in the literature that sentiment analysis refers to identifying whether someone expresses a positive or negative attitude (Yadollahi et al., 2017). Sentiments are differentiated from emotions by the duration in which they are experienced (Munezero et al., 2014). In contrast, EA is primarily interested in specific emotions such as anger or joy. In the literature reviewed, the task of emotion recognition/detection/classification/prediction refers to the goal of identifying a specific emotion expressed by the author of a particular utterance or set of utterances (Mohammad et al., 2018; Chatterjee et al., 2019). In some cases, it refers to the task of identifying the cause of emotions in a broader context, for example, how speakers affect each other’s emotional state in a conversation (Liu et al., 2022; Zhan et al., 2022). We discuss in Section 4.2 the issues that arise from the lack of consistent terminology.

3.2. Emotion models and frameworks

Several conceptual models from psychologists and neuroscientists have emerged to categorize and interpret emotions. Notably, three prominent models stand out: the discrete or categorical model, the dimensional model, and the componential model. They are described in the following:

- **Discrete or categorical model**: Emotions are categorized into distinct classes and categories in this model. Each emotion is treated as independent and separate from others. This model includes well-known frameworks such as Ekman’s model (Ekman and Friesen, 1981) consisting of six basic emotions (anger, fear, sadness, joy, disgust and surprise) and Plutchik’s model (Plutchik, 1982), which encompasses eight primary emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust).

- **Dimensional model**: Unlike the previous model, dimensional models view emotions as interconnected rather than independent entities. This model acknowledges that emotions can vary in intensity and can be represented along different dimensions. For instance, emotion names are located in vector spaces of affect that consider valence, arousal, and dominance (VAD) (Russell and Mehrabian, 1977).

- **Componential model**: The componential model comprised the appraisal theory (Scherer, 1999; Lazarus, 1991), which emphasizes that emotions are influenced by a person’s subjective evaluation or appraisal of a situation or experience. Factors such as responsibility, certainty, pleasantness, control, and attention play a significant role in influencing an individual’s emotional responses to a particular situation or experience (Smith and Ellsworth, 1985).

Our analysis reveals a clear prevalence of discrete conceptual models in the literature, indicating a strong interest in classifying emotions into distinct categories and treating each emotion as an independent entity – in line with standard supervised learning objectives. However, a few papers use dimensional models, and an even smaller proportion explore componential models.

Research papers in discrete or categorical emotion models commonly rely upon different theories.
The most influential theory is the Ekman model (Ekman and Friesen, 1981), followed by the Plutchik model (Plutchik, 1982). A few other studies considered the Turne list (Turner, 2000) and the Ortony, Clore, and Collins model (Ortony et al., 1988). With dimensional models, some studies used a mapping approach to associate emotion categories with dimensions. For instance, Bagher Zadeh et al. (2018) conducted data annotating mapping the Ekman emotions to a Likert scale in which values are assigned from 0 to 3 to represent the presence of emotions. Park et al. (2021) introduced a framework that leverages VAD scores obtained from an emotion lexicon to acquire VAD scores from sentences labeled with categorical emotions. More recently, component models have gained attention and been introduced in the field. Hofmann et al. (2020) established an annotation task where event descriptions are linked to seven appraisal dimensions (Smith and Ellsworth, 1985). Stranisci et al. (2022) relied on Roseman's model of appraisal (Roseman, 1991) to annotate a corpus of social media posts. More recently, Troiano et al. (2023) built a corpus where people described emotion-triggering events and their appraisals. Subsequently, readers were asked to reconstruct emotions and appraisals from the text. They provide a thorough appraisal theory overview for text-based emotion analysis suitability assessment. Additionally, it shows that appraisals offer an alternative computational approach to EA and enhance emotion categorization in text using joint models.

Figure 4 provides an overview of the most frequent emotion categories in the literature. Due to the widespread use of Ekman's model (Ekman and Friesen, 1981), the prevalent emotions observed are anger, fear, sadness, joy, disgust, and surprise. Additionally, Plutchik's model, which is the second most used, highlights the frequency of trust and anticipation. However, a range of other emotions receives comparatively less attention. These include excitement, anxiety, optimism, pessimism, guilt, shame, and love.

3.3. Data

In this section, we provide an overview of the datasets that have been extensively used in various EA studies and more recently developed datasets. In our analysis, we identified over 50 papers that created a dataset. Table 1 gives a summary list of some of these datasets, showcasing their key characteristics: annotation scheme, language, multimodality, source, and size.

Most of the datasets found in the EA studies are focused on emotion recognition in conversation, including IdeDialog (Ide and Kawahara, 2022), M3ED (Zhao et al., 2022), MuSE (Jaiswal et al.,
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Annotation</th>
<th>Lang.</th>
<th>Source</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>IdeDialog (Ide and Kawahara, 2022)</td>
<td>P</td>
<td>JA</td>
<td>Twitter dialogues</td>
<td>13,806</td>
</tr>
<tr>
<td>M3ED (Zhao et al., 2022)</td>
<td>E + neutral</td>
<td>CN</td>
<td>TV dialogues</td>
<td>24,449</td>
</tr>
<tr>
<td>MuSE (Jaiswal et al., 2020)</td>
<td>sad, anger, contentment, amusement, neutral</td>
<td>EN</td>
<td>Recordings</td>
<td>28 speakers</td>
</tr>
<tr>
<td>MELD (Zahiri and Choi, 2017; Poria et al., 2019)</td>
<td>E + neutral</td>
<td>EN</td>
<td>Movie dialogues</td>
<td>13,000</td>
</tr>
<tr>
<td>DailyDialog (Li et al., 2017)</td>
<td>E + other</td>
<td>EN</td>
<td>Movie dialogues</td>
<td>13,118</td>
</tr>
<tr>
<td>EmoryNLP (Zahiri and Choi, 2017)</td>
<td>E + neutral</td>
<td>EN</td>
<td>TV show dialogues</td>
<td>12,606</td>
</tr>
<tr>
<td>IEMOCAP (Busso et al., 2008)</td>
<td>happiness, anger, sadness, frustration, neutral</td>
<td>EN</td>
<td>Dyadic interactions</td>
<td>10,039</td>
</tr>
<tr>
<td>StudEmo (Ngo et al., 2022)</td>
<td>28 emotion categories</td>
<td>EN</td>
<td>Customer reviews</td>
<td>5,182</td>
</tr>
<tr>
<td>RED v2 (Ciobotaru et al., 2022)</td>
<td>P - disgust, anticipation</td>
<td>RO</td>
<td>Tweets</td>
<td>5,449</td>
</tr>
<tr>
<td>APPReddit (Stranisci et al., 2022)</td>
<td>appraisals (Roseman, 1991)</td>
<td>EN</td>
<td>Reddit</td>
<td>1,091</td>
</tr>
<tr>
<td>Universal Joy (Lamprinidis et al., 2021)</td>
<td>anger, anticipation, fear, joy, sadness</td>
<td>EN</td>
<td>Facebook posts</td>
<td>530k</td>
</tr>
<tr>
<td>WRIME (Kajiwara et al., 2021)</td>
<td>P, emotion intensity</td>
<td>JA</td>
<td>SNS posts</td>
<td>17,000</td>
</tr>
<tr>
<td>EmoEvent (Plaza del Arco et al., 2020)</td>
<td>E + neutral</td>
<td>EN-ES</td>
<td>Tweets</td>
<td>8,409</td>
</tr>
<tr>
<td>GoodNewsEveryone (Bostan et al., 2020)</td>
<td>15 emotion categories</td>
<td>EN</td>
<td>News Headlines</td>
<td>5,000</td>
</tr>
<tr>
<td>PO-EMO (Haider et al., 2020)</td>
<td>joy, sadness, uneasiness, vitality, suspense, sublime, humor, annoyance, nostalgia</td>
<td>EN-DE</td>
<td>Poems</td>
<td>64 (EN), 18 (DE)</td>
</tr>
<tr>
<td>enISEAR (Hofmann et al., 2020)</td>
<td>appraisals (Smith and Ellsworth, 1985)</td>
<td>EN</td>
<td>Self-reports</td>
<td>1,001</td>
</tr>
</tbody>
</table>

Table 1: EA datasets. [E] Ekman, [P] Plutchik. The first part of the table contains multimodal datasets used for emotion recognition in conversation. The second part contains datasets used in EA and recent ones. Lang.: Language

In the second part of the table, we summarize some of the recent EA datasets created over the past few years. These datasets include StudEmo (Ngo et al., 2022), RED v2 (Ciobotaru et al., 2022), APPReddit (Stranisci et al., 2022), Universal Joy (Lamprinidis et al., 2021), WRIME (Kajiwara et al., 2021), EmoEvent (Plaza del Arco et al., 2020), GoodNewsEveryone (Bostan et al., 2020), PO-EMO (Haider et al., 2020), and enISEAR (Hofmann et al., 2020). The annotation schemes of these datasets offer a wider range of emotion frameworks and emotions compared to the previously described for emotion recognition in conversation. While most adhere to the discrete emotion model, enISEAR and APPReddit incorporate the componential model with appraisals. Although English is the most predominant language, multilingual datasets like EmoEvent and Universal Joy encompass multiple languages. RED includes Romanian, and WRIME includes Japanese, further diversifying the linguistic coverage of emotion recognition studies. None of these datasets present a mul-
timodal aspect. The instances are sourced from diverse domains, including social media (Twitter, Facebook), news headlines, self-reports, poems, customer reviews, and community networks (Reddit). Among these domains, social media emerges as the most predominant data source.

Despite the highly subjective nature of emotions, only a few works (Martin et al., 2022a; Zhan et al., 2022; Ghosh et al., 2022) mentioned the gathering of annotator demographic information, yet most of them do not use this information for their methods.

### 3.4. Applications

Studying patterns of human emotions and understanding people’s feelings are essential in many real-life applications (Picard, 2000). In our analysis, among the most used applications, we found that emotion recognition in conversation has gained significant popularity in the field by analyzing the emotions expressed during interactions (Hazarika et al., 2018; Zhong et al., 2019; Ghosal et al., 2020; Hu et al., 2021; Lee and Lee, 2022). Another noteworthy application within the realm of EA is its application in the analysis of narrative texts (Kim and Klinger, 2018b; Liu et al., 2019; Haider et al., 2020; Zad and Finlayson, 2020; Cortal et al., 2023). Additionally, EA has found a valuable niche in the healthcare sector, particularly in addressing mental health concerns (Khanpour and Caragea, 2018). With the advent of social media, detecting and monitoring emotional states associated with mental well-being has become possible. This capability has proven invaluable in identifying individuals at risk of self-harm, anxiety, or stress, as detailed by Turcan et al. (2021). EA also proves beneficial during critical events like the COVID-19 pandemic (Ng et al., 2020; Sosea et al., 2022; Sampath et al., 2022). Beyond these applications, EA extends its impact to various other domains including email customer care (Gupta et al., 2010), span prediction (Alhuzali and Ananiadou, 2021), fake news (Meleng et al., 2019), moral content detection (Asprino et al., 2022), sarcasm detection (Chauhan et al., 2020a,b), time series analysis of emotional loading in central bank statements (Buechel et al., 2019), and financial forecasting (Seroyizhko et al., 2022).

### 4. Shortcomings of EA in NLP

In this section, we outline the main areas where there is room for improvement in EA research.

#### 4.1. Lack of Diversity in Available Datasets

We find significant overlap between datasets: most are in English, sourced from movies or TV shows, and annotated using Ekman’s emotions categorically. The lack of language diversity in NLP has been discussed at length (Martin et al., 2022b; Sonu et al., 2022), and it is not specific to EA, but EA lacks diversity in other aspects:

##### Diversity in emotion labels.

Section 3.3 shows that most datasets are labeled according to Ekman’s basic emotions, or Plutchik to a lesser extent. While these emotions are widely used in the analyzed studies, they do not generalize to all tasks or domains (De Bruyne, 2023). Human emotions are nuanced because this nuance is required for understanding the world. Yet current datasets predominantly focus on a few coarse-grained emotions. For example, feeling guilty or lonely both fall under sadness but convey very different meanings. Furthermore, the most commonly annotated emotions are negative, particularly anger, fear, and sadness. This emphasis on negative emotions has far-reaching consequences for EA and downstream applications. Detecting more fine-grained negative emotions might be helpful in a mental disorder detection setting, but positive emotions may be more relevant for tutoring systems as they aid learning. A notable example demonstrating the importance of aligning emotion categories with the specific task at hand is the work of (Haider et al., 2020), where aesthetic emotions play a pivotal role in analyzing emotions within poems. This approach encompasses emotions such as joy, sadness, uneasiness, vitality, suspense, sublime, humor, annoyance, and nostalgia. It is also noteworthy that most of the papers in the survey did not specify any concrete application of EA—lacking guidance in selecting the labels that will aid in a given task.

##### Diversity in annotation schemes.

The preference for Ekman’s or Plutchik’s models significantly limits the adaptability of EA techniques to different contexts and tasks (Plaza-del Arco et al., 2022). Each NLP task may require a nuanced understanding of emotions that these models might not fully capture. This limitation impedes the development of more versatile and context-specific emotion analysis approaches. Furthermore, this preference for specific models neglects the rich landscape of existing psychological theories related to emotions. Emotion is a complex and multifaceted phenomenon, and various psychological theories offer different insights into how emotions are generated, expressed, and perceived (Hofmann et al., 2020). Overlooking these diverse theories means missing out on knowledge, valuable perspectives, and dimensions of emotion. More recent studies follow alternative theories of cognitive appraisal of events (see Section 3.2) and show their potential...
for emotion classification when encoded in categorical models. Öhman (2020) also point out that current annotation schemes are based on psychological theories not primarily focused on text.

4.2. Inconsistency in Terminology

Section 3.1 shows the inconsistent terminology used in EA. This lack of specificity can obscure the related but distinct subtasks. Currently, no distinction is made between ‘emotion classification’, for example, and ‘emotion prediction’. We generally understand classification as a subtask of prediction, so these terms might be treated as synonymous. We argue that EA should clarify its jargon. Given the growing interest in detecting the triggers for emotion in NLP, we suggest using the term prediction to demarcate tasks that are interested in emotions before they occur, such as forecasting a customer’s frustration during a support chat based on contextual cues. One could also make the case that emotion classification should refer to classifying a set of emotions, whereas emotion detection or recognition are searching for a specific emotion for a specific reason. For example, using NLP for suicide prevention might involve detecting sadness, desperation, or anger. However, it might seem odd to say it would involve classifying emotions. Mapping EA in NLP using consistent nomenclature will enable NLP practitioners to focus on fine-grained specific tasks since there will be a clear goal. It will also highlight gaps in EA for which new tasks must emerge.

4.3. Lack of Interdisciplinarity

Two issues arise out of the general lack of interdisciplinary engagement in EA: The first issue was already raised by Kusal et al. (2022a), who concluded that it would be helpful to gain a deeper understanding of emotions for the classification process. As we showed in Section 3.2, currently, EA in NLP relies solely on emotion models and frameworks borrowed from psychology. In particular, NLP overwhelmingly relies on Ekman’s theory from the 1970s. This theory identifies emotions with bodily signatures, while Barrett (2017) has empirically problematized it by stating that emotions do not have such signatures and emphasizes the importance of considering the context for interpreting emotions. Additionally, Ekman’s views have been conceptually criticized by philosophers who argue that emotions possess intentionality, being directed towards something, a characteristic not adequately addressed in Ekman’s theory. 6

The second issue concerns potential applications across fields. Most papers we reviewed here do not mention the potential applicability of EA to the humanities and social sciences, even though that is a fruitful area of development (e.g., digital humanities). Collaboration between these disciplines can lead to innovative research, the development of specialized tools, and the integration of emotional insights into a broader range of academic studies. This interdisciplinary synergy would be mutually beneficial, enriching the understanding of human emotions and their role in various domains.

4.4. Demographics and Cultural Implications

Demographic factors such as age, gender, cultural background, and socioeconomic status can significantly shape individual differences while expressing and experiencing emotions. Bender and Friedman (2018) call for careful data curation and conscientious reporting of the processes and actors involved in corpus creation. More specifically, they call for reporting who produced and who annotated the data. We note that most papers we reviewed do not include a data statement or report demographics for either the data creators or the annotators. Additionally, when demographic data is collected, it is more commonly associated with annotators rather than data creators, as observed in prior studies ((Troiano et al., 2019; Haider et al., 2020; Kajiwara et al., 2021). However, it is noteworthy that in most cases, this demographic information is not integrated into the model inputs. Given the subjective nature of emotions and the enormous role annotators and sources can play, this observation highlights a concerning gap in the field. It also emphasizes the need to consider the influence of annotator demographics for a more comprehensive and nuanced approach to EA. Furthermore, recent work (Basile et al., 2021; Plank, 2022; Davani et al., 2022) has suggested leveraging human label variation when annotating and creating data. Subjective tasks like EA would benefit from this approach as it allows for developing more inclusive and diverse systems considering individual perspectives and using annotation to explore the range of possible emotions in a task (the descriptive paradigm of Rottger et al., 2022).

5. Discussion

We can now answer the four questions from the beginning of our study:

1. How are EA tasks defined in NLP?

The commonly used terms to refer to EA tasks include emotion detection/classification/recognition/prediction.
being emotion recognition the most used. These terms are often used interchangeably in different research papers. Consequently, this interchangeable usage lacks specificity and can potentially obscure the distinct, yet interconnected, subtasks involved.

2. What are the most prominent emotion frameworks and which emotions are modeled? The majority of studies primarily rely on Ekman’s basic emotions, with Plutchik’s model being used to a lesser extent. Consequently, other psychological theories are ignored to some degree. This limits the usefulness of EA models in downstream applications. Recent studies based on cognitive appraisals show future directions to encode categorical models. Future work should endeavor to diversify the emotions and emotions frameworks represented.

3. Is the subjectivity of emotions considered in terms of demographics and cultural factors? No consideration is given to the subjectivity of emotions in relation to demographics and cultural factors. The subjective nature of emotions in EA restricts the diversity and inclusivity of NLP methods. Therefore, it is crucial for future endeavors in dataset creation and NLP system development to take these factors into account.

4. What are the primary NLP applications for EA? Among the papers reviewed, the majority of them focus on emotion recognition in conversation which involves the analysis of emotions expressed during interactions and generating empathetic responses. Additionally, there are applications in areas such as public health, critical event analysis, email customer care, and financial forecasting.

6. Roadmap for Future Directions

Demographics for Diversity and Inclusivity. Collecting data on the demographics of annotators or data creators, such as age, gender, ethnicity, and cultural background, provides insights into how individual perspectives shape emotional experiences. Integrating demographic information into the NLP model inputs could lead to more context-aware and tailored EA. Research has shown that our perception of emotions is strongly gendered, racialized, and age-dependent, and datasets should capture this diversity. Moreover, while demographics can affect how emotions are perceived, individual differences also impact this (Orlikowski et al., 2023).

Avoiding the aggregation of labels and instead focusing on individual demographics promotes a more nuanced and inclusive approach to EA.8

Tailor Emotion Categories to your Task. When choosing emotion models or categories, it is vital to not solely depend on the widely repeated theory in the literature. Instead, one should prioritize considering the specific domain and application scenario. In addition, to avoid fixed categorization and in line with standard supervised learning objectives, exploring models that can transfer knowledge across different emotions is desirable. For instance, recent zero or few-shot learning paradigms offer promising alternatives that enable more flexible and adaptable approaches in the field of EA (Plaza-del Arco et al., 2022).

EA Nomenclature. EA is the broader area that comprises EA tasks in NLP. ‘Emotion detection’, ‘emotion classification’, and ‘emotion recognition’ are often used interchangeably as synonyms. However, it can be beneficial to distinguish between these terms based on the specific task at hand. For example, using NLP for mental disorder detection might involve detecting specific emotions as opposed to classifying the emotions involved. In addition, other terms should not be used as synonyms, given that they contribute to a lack of clarity about the task’s aims (e.g., ‘prediction’ and ‘classification’).

Interdisciplinarity. EA is inherently interdisciplinary and borrows from social science and humanities. EA should keep up to date with current theories and models of emotion (e.g., the discrete vs. dimensional vs. componential models). Furthermore, given that NLP’s domain is language, it would stand to benefit from a more pluralistic understanding of emotion engaging with philosophical theories of emotion, such as so-called cognitive theories of emotion and perceptual theories of emotion.9

7. Conclusion

EA has grown in NLP since the 2010s, with more research on human emotions in numerous fields. Over time, conferences and publications have published various studies on the area, demonstrating its growing popularity. We reviewed over 150 ACL anthology papers from 2014 to 2022 in this report. We discussed

8See the Perspectivist Data initiative https://pdai.info/
9For a survey of theories of emotion, see Brady (2018).
field nomenclature, emotion theories, demographic and cultural subjectivity, and NLP applications.

We found demographic and cultural gaps, inadequate emotion category match to the downstream goal, no standard systematic nomenclature in EA, and no interdisciplinary research. For each gap, we proposed future directions to develop meaningful linkages, facilitate targeted study, and enable NLP emotion modeling for nuance.

Limitations

Our survey primarily focuses on papers included in the ACL anthology. However, we acknowledge the existence of other relevant EA papers in the NLP field that may have been published in journals or other conferences. Furthermore, we are aware that there are papers that may employ varying keywords or terminology when describing their work related to EA.

Finally, while we have conducted a thorough analysis of specific aspects related to EA, there are other facets that we have not covered in this review.

Ethics Statement

The paper is related to a survey, therefore, it does not involve ethical issues in the methodology.

Acknowledgements

The work of Flor Miriam Plaza-del-Arco, Amanda Cercas Curry and Dirk Hovy has been funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR). They are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

8. Bibliographical References


Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. 2021. Universal joy a data set and results for classifying emotions across languages. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages


Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotion’s in Social Media. Association for Computational Linguistics, Barcelona, Spain (Online).


Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modelling human label variation goes beyond sociodemographics.


